LIMITED ERROR BASED EVENT LOCALIZING TEMPORAL DECOMPOSITION*

Phu Chien Nguyen, Masato Akagi
Graduate School of Information Science
Japan Advanced Institute of Science and Technology
1-1 Asahidai, Tatsunokuchi, Ishikawa 923-1292, Japan
e-mail: {chien, akagi}@jaist.ac.jp

ABSTRACT

This paper proposes a low delay temporal decomposition (TD) method for line spectral frequency (LSF) parameters called "*Limited Error Based Event Localizing Temporal Decomposition*" (LEBEL-TD, for short). In previous work with TD, TD analysis was usually performed on each speech segment of about 200-300 ms or more, making it impractical for online applications. In this present work, the event localization is determined based on a limited error criterion and a local optimization strategy, which results in an average algorithmic delay of 75 ms. Simulation results show that an average log spectral distortion of about 1.5 dB can be achievable at an event rate of 20 events/sec. Also, LEBEL-TD uses neither the computationally costly singular value decomposition routine nor the event refinement process, thus reducing significantly the computational cost of TD. It is shown that excitation information of speech can be well described using the LEBEL-TD technique.

1 INTRODUCTION

Temporal decomposition (TD) of speech initiated by Atal [1] involves the decomposition of a sequence of spectral parameter vectors, i.e. linear predictive coding (LPC) parameters, into a series of overlapping event functions and an associated series of event vectors as given in Equation (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 vector and 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 and K are the number of frames and number of events in the block of spectral parameters under consideration, respectively.

The second order TD model used in [2], where, only two adjacent event functions overlap, is given in Equation (2).

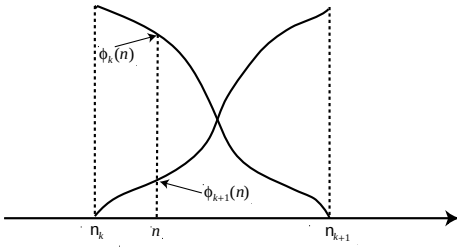

Fig. 1: Example of two adjacent event functions in the second order TD model.

$$\hat{\mathbf{y}}(n) = \mathbf{a}_k \phi_k(n) + \mathbf{a}_{k+1} \phi_{k+1}(n), \quad n_k \le n < n_{k+1} \tag{2}$$

where, n_k and n_{k+1} are the locations of event k and event $k+1$, respectively.

The restricted second order TD model was utilized in [3, 4, 6] with an additional restriction to the event functions in the second order TD model is that all event functions at any time sum up to one. The argument for imposing this constraint on the event functions can be found in [3]. Equation (2) is rewritten as

$$\hat{\mathbf{y}}(n) = \mathbf{a}_k \phi_k(n) + \mathbf{a}_{k+1} \left(1 - \phi_k(n)\right), \quad n_k \le n < n_{k+1} \tag{3}$$

Despite the fact that TD has the potential to become a versatile tool in speech analysis, its high computational complexity and long algorithmic delay make it impractical for online applications. In the original TD method by Atal [1], TD analysis was performed on each speech segment of about 200-300 ms, thus resulting in an algorithmic delay of more than 200 ms. In addition, this method is very computationally costly, which has been mainly attributed to the use of the singular value decomposition (SVD) routine and the iterative refinement process. These prevent the Atal's method from online applications. The method proposed in [5] for TD, S^{2}BEL-TD, reduced the computational cost of TD by avoiding the use of SVD, but the long algorithmic delay has more or less remained the same. S^{2}BEL-TD uses a spectral stability criterion to determine the initial event locations. Meanwhile, the event localization in the optimized TD (OTD) method [2] was performed

*This work was supported by CREST (Core Research for Evolutional Science and Technology) of JST.

using an optimized approach (dynamic programming). The OTD method can achieve very good results in terms of reconstruction accuracy, but its long algorithmic delay (more than 450 ms) makes it suitable for speech storage related applications only. Also, both the OTD and S^{2}BEL-TD methods use the line spectral frequency (LSF) parameters as input, which might cause the corresponding LPC synthesis filter to be unstable. The restricted TD (RTD) [4] and the modified RTD (MRTD) [6] methods considered the ordering property of LSFs to make LSF parameters possible for TD. These methods require an average algorithmic delay of about 95 ms, while can achieve relatively good results.

In this paper we propose a new algorithm for temporal decomposition of LSF parameters called "*Limited Error Based Event Localizing Temporal Decomposition*" (LEBEL-TD). This method employs the restricted second order model and a novel approach to event localization. Here, the event localization is initially performed based on a limited error criterion, and then further refined by a local optimization strategy. In the following, the event vectors are set as the original LSF parameters at the event locations and thus, the ordering property of LSFs does not need to be considered. This algorithm for TD requires only 75 ms average algorithmic delay, while can achieve results comparable to the S^{2}BEL-TD, RTD and MRTD methods. Moreover, LEBEL-TD uses neither the computationally costly SVD routine nor the iterative refinement process, thus resulting in a very low computational cost required for TD analysis.

2 LEBEL-TD OF SPEECH

2.1 Determination of Event Functions

Assume that the locations n_k and n_{k+1} of two consecutive events are known. Then, the right half of the k^{th} event function and left half of the $(k+1)^{th}$ event function can be optimally evaluated by using $a_k = y(n_k)$ and $a_{k+1} = y(n_{k+1})$. The reconstruction error, $E(n)$, for the n^{th} spectral parameter vector is

$$\begin{aligned} E(n) &= \| y(n) - \hat{y}(n) \|^2 \\ &= \| (y(n) - a_{k+1}) - (a_k - a_{k+1})\phi_k(n) \|^2 \end{aligned} \quad (4)$$

where, $n_k \leq n < n_{k+1}$. Therefore, $\phi_k(n)$ should be determined so that $E(n)$ is minimized.

We suggest using the determination of event functions, where, geometrically speaking, the two event vectors $\mathbf{a}_k$ and $\mathbf{a}_{k+1}$ define a plane in the spectral parameter vector space. The determination of event functions is depicted in Fig. 2 as the projection of vector $\mathbf{y}(n)$ onto this plane. In order to make all event functions well-shaped and model the temporal structure of speech more effectively, the event functions are determined corresponding to the point, $\hat{\mathbf{y}}(n)$, of the line segment between $\hat{\mathbf{y}}(n-1)$ and $\mathbf{a}_{k+1}$ with minimum distance from $\mathbf{y}(n)$. Here, by a well-shaped event function we mean an event function having a growth cycle; during which the event function

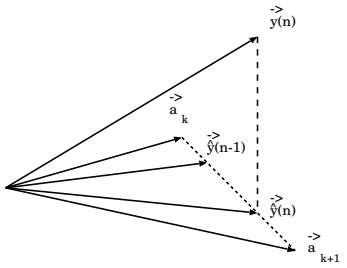

Fig. 2: Determination of the event functions in the transition interval $[n_k, n_{k+1}]$.

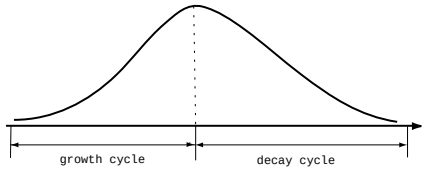

Fig. 3: Example of a well-shaped event function

grows from zero to one, and a decay cycle; during which the event function decays from one to zero, as shown in Fig. 3. This helps to prevent the event functions from having more than one lobe, which is not acceptable in the conventional TD. Also, this reduces the quantization error of event functions when vector quantized.

In mathematical form, the above determination of event functions can be written as

$$\phi_k(n) = \begin{cases} 1 - \phi_{k-1}(n), & \text{if } n_{k-1} < n < n_k \\ 1, & \text{if } n = n_k \\ \min(\phi_k(n-1), & \\ \quad \max(0, \hat{\phi}_k(n))), & \text{if } n_k < n < n_{k+1} \\ 0, & \text{otherwise} \end{cases} \quad (5)$$

where,

$$\hat{\phi}_k(n) = \frac{\langle (\mathbf{y}(n) - \mathbf{a}_{k+1}), (\mathbf{a}_k - \mathbf{a}_{k+1}) \rangle}{\| \mathbf{a}_k - \mathbf{a}_{k+1} \|^2} \quad (6)$$

2.2 LEBEL-TD Algorithm

The section of spectral parameters, $y(n)$, where $n_k \leq n < n_{k+1}$, is termed a segment. The total accumulated error, $E_{seg}(n_k, n_{k+1})$, for the segment is

$$E_{seg}(n_k, n_{k+1}) = \sum_{n=n_k}^{n_{k+1}-1} E(n) \quad (7)$$

where, $E(n)$ can be calculated for every $n_k \leq n < n_{k+1}$ once n_k and n_{k+1} are known. The buffering technique for LEBEL-TD is depicted in Fig. 4, and the whole algorithm is described as follows.

Step 0. Set $k \leftarrow 1$, $n_1 \leftarrow 1$, $a_1 \leftarrow y(1)$; set n_2 as the last location from n_1 on so that the reconstruction error for every frame in the interval (n_1, n_2) is less than a predetermined number ϵ.

Step 1. Similarly, set n_3 as the last location from n_2 on

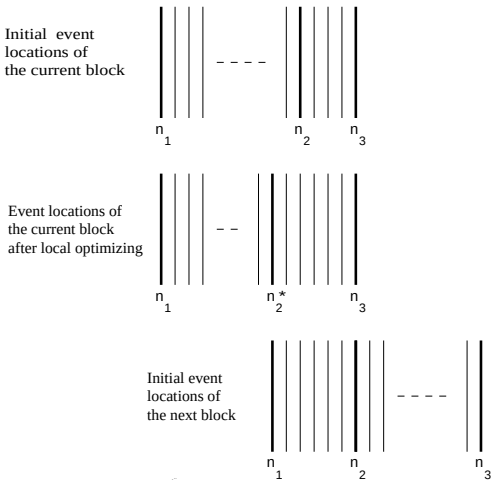

Fig. 4: Buffering technique for LEBEL-TD

so that the reconstruction error for every frame in the interval (n_2, n_3) is less than ϵ.

Step 2. Local optimize the location of n_2 in the interval (n_1, n_3).

$$n_2^* = \arg \min_{n_1 < n_2 < n_3} \{E_{seg}(n_1, n_2) + E_{seg}(n_2, n_3)\}$$

where, only n_2 that makes $E(n) < \epsilon$ for every $n_1 < n < n_3$ is taken into account. If n_3 is the last frame, set $k \leftarrow k+1$, $a_k \leftarrow y(n_2^*)$, $a_{k+1} \leftarrow y(n_3)$; and exit.

Step 3. Set $k \leftarrow k+1$, $a_k \leftarrow y(n_2^*)$; then set $n_1 \leftarrow n_2^*$, $n_2 \leftarrow n_3$; and go back to step 1.

The predetermined number ϵ is called the reconstruction error threshold, and it is the only parameter that effects the number and locations of the events. It controls the event rate, i.e. the number of events per second, and can be appropriately selected to achieve the optimal performance of LEBEL-TD for different applications.

In the LEBEL-TD method, the event vectors are set as the spectral parameter vectors corresponding to the event locations. Obviously, the event vectors are valid LSF vectors and the stability of the corresponding LPC synthesis filter can be thus ensured after spectral transformation performed by LEBEL-TD.

3 PERFORMANCE EVALUATION

Objective performance evaluation was performed based on the log spectral distortion (LSD) between the original LSF parameters, $y(n)$, and the reconstructed LSF parameters, $\hat{y}(n)$ [7].

A set of 250 sentences of the ATR Japanese speech database were selected as the speech data. This speech data set consists of about 20 minutes of speech spoken by 10 speakers (5 male & 5 female) re-sampled at 8 kHz sampling frequency. 10^{th} order LSF parameters were calculated using a LPC analysis window of 30 ms at 10 ms frame intervals, and TD analyzed using the LEBEL-TD method. Here, $\epsilon = 0.045$ was empirically chosen as a suitable value for the reconstruction error threshold. Fig. 5 shows the plot of event functions obtained from the LEBEL-TD method for an example of a Female/Japanese speech sentence.

Table I: Event rate, average LSD, and percentage number of outlier frames obtained from the LEBEL-TD, S^{2}BEL-TD, RTD, and MRTD methods.

Method	Event rate	Avg. LSD	> 4 dB
LEBEL-TD	19.996	1.512 dB	0.07%
S^{2}BEL-TD	19.455	1.464 dB	0.94%
RTD	20.163	1.563 dB	0.96%
MRTD	20.163	1.568 dB	0.98%

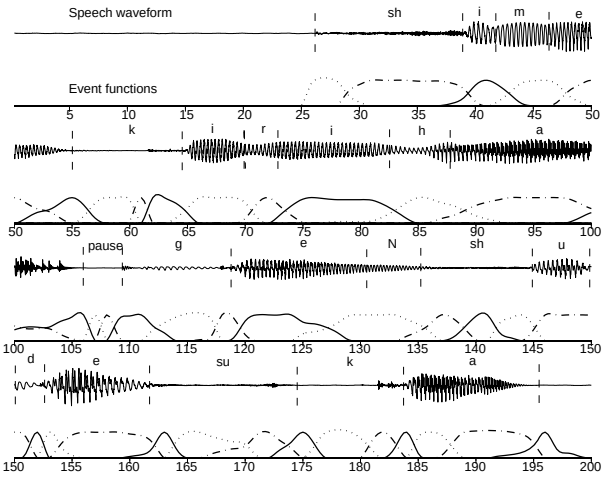

Fig. 5: Plot of the event functions obtained from the LEBEL-TD method for the Female/Japanese speech sentence "*shimekiri ha geNshu desu ka*". The speech waveform is also shown together with the phonetic transcription for reference.

Table I gives the summary of LSD and the event rate obtained from LEBEL-TD for the above set of sentences. For comparison, the results obtained from the S^{2}BEL-TD, RTD and MRTD methods for the same speech data set are also shown in this table. Results indicate slightly better performance in case of S^{2}BEL-TD and LEBEL-TD over RTD and MRTD. Note that the S^{2}BEL-TD and RTD methods, however, have not ensured the stability of the corresponding LPC synthesis filter [6].

Also, the event rate was found to be about 20 events/sec, thus resulting in an average buffering delay of about 50 ms, i.e. 5 frames, along with a 10 ms, i.e. one frame, look-ahead. Considering that the LPC analysis window is 30 ms long, which implies a 15 ms look-ahead. Therefore, the average algorithmic delay for LEBEL-TD is about 75 ms and has been known to be the lowest algorithmic delay for TD so far. Moreover, LEBEL-TD has significantly reduced the computational cost of TD because it uses neither the computationally costly SVD routine nor the iterative refinement process.

Table II: Event rate, average LSD, and percentage number of outlier frames obtained from the LEBEL-TD method for some ϵ.

ϵ	Event rate	Avg. LSD	2-4 dB	>4 dB
0.031	25.028	1.233 dB	18.69%	0.00%
0.045	19.996	1.512 dB	32.52%	0.07%
0.072	15.059	1.922 dB	48.52%	1.60%

These make LEBEL-TD suitable for online applications.

We have also evaluated the performance of LEBEL-TD on the above speech data set for some ϵ. Table II gives the summary of LSD and the event rate obtained from LEBEL-TD for different values of ϵ. As can be seen from the table, the event rate decreases and the average LSD increases as ϵ increases.

4 LEBEL-TD OF SPEECH EXCITATION

We employ the technique used in [5] to describe the temporal characteristic of the speech excitation parameters. According to [5], the same event functions obtained from LEBEL-TD analysis of LSF parameters are also used to describe the temporal evolution of the gain, pitch, and voicing parameters. The so called excitation targets are estimated from the excitation parameters and the event functions in the least mean square sense. In the case of pitch parameters, linear interpolation was used within the unvoiced segments to form a continuous contour. In the case of voicing parameters, a hard limiter with a threshold value of 0.5 was used to determine the binary voicing targets and reconstructed binary voicing parameters from the non-binary voicing results.

The performance of LEBEL-TD in terms of excitation parameters has also been evaluated over the speech data set used in Section 3. The RMS gain error, RMS pitch error and percentage number of frames with voicing errors obtained from LEBEL-TD analysis of excitation parameters were found to be about 3.42 dB, 6.66 Hz, and 4.45%, respectively. Meanwhile, the corresponding results obtained from the S^{2}BEL-TD method for this speech data set were found to be about 4.27 dB, 6.25 Hz, and 5.16%. 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. Also, no voicing errors were observed during continuous voiced and unvoiced segments, except for the points of voicing transition.

5 CONCLUSION

In this paper we have presented a low delay method for temporal decomposition of LSF parameters. The proposed LEBEL-TD method uses the limited error criterion for initially estimating the event locations, and then further refines them using the local optimization strategy. This method achieves results comparable to other TD methods such as S^{2}BEL-TD, RTD, MRTD

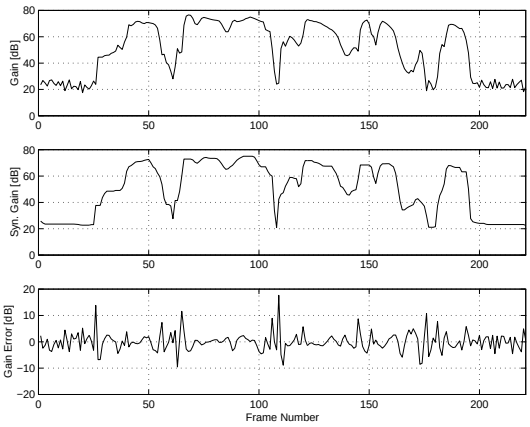

Fig. 6: Original and reconstructed gain contours, and the gain error for the same speech sentence as in Fig. 5. The RMS gain error is 3.4 dB.

while requiring less algorithmic delay and less computational cost. Moreover, the buffering technique used for continuous speech analysis has been well developed and the stability of the corresponding LPC synthesis filter after spectral transformation performed by LEBEL-TD has been completely ensured. It is shown that the temporal pattern of the speech excitation parameters can also be well described using the LEBEL-TD technique.

6 REFERENCES

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

[2] C.N. Athaudage, A.B. Brabley and M. Lech, "Optimization of a temporal decomposition model of speech", *Proc. ISSPA'99*, pp. 471-474, 1999.

[3] P.J. Dix and G. Bloothooft, "A breakpoint analysis procedure based on temporal decomposition", *IEEE Trans. on Speech and Audio Processing*, Vol. 2, No. 1, pp. 9-17, 1994.

[4] S.J. Kim, S.H. Lee, W.J. Han and Y.H. Oh, "Efficient quantization of LSF parameters based on temporal decomposition", *Proc. ICSLP'98*, pp. 2575-2578, 1998.

[5] A.C.R. Nandasena, P.C. Nguyen and M. Akagi, "Spectral stability based event localizing temporal decomposition", *Computer Speech and Language*, Vol. 15, No. 4, pp. 381-401, 2001.

[6] P.C. Nguyen and M. Akagi, "Improvement of the restricted temporal decomposition method for line spectral frequency parameters", *Proc. ICASSP'02*, pp. 265-268, 2002.

[7] K.K. Paliwal, "Interpolation properties of linear prediction parametric representations", *Proc. Eurospeech'95*, pp. 1029-1032, 1995.