Temporal decomposition of parameter tracks for speech coding

Miguel Arjona Ramírez
University of São Paulo
Signal Processing Laboratory, PSI, EPUSP
Av. Prof. Luciano Gualberto, trav. 3, 158 - 05508-970
São Paulo - SP - Brazil
miguel@lps.usp.br

Vinícius Oliveira Pinheiro Machado
University of São Paulo
Signal Processing Laboratory, PSI, EPUSP
Av. Prof. Luciano Gualberto, trav. 3, 158 - 05508-970
São Paulo - SP - Brazil
vinicius.machado@poli.usp.br

***Abstract*— Speech coders often interpolate parameters extracted at regular intervals. Temporal decomposition (TD) provides means for locating significant parameter changes at instants when paramater values are extracted as target vectors. The original TD algorithm based on singular value decomposition (SVD) is adapted for line spectral frequency (LSF) tracks and a criterion is proposed for setting the number of refinement iterations. Greater control over the shape and extent of the event functions is provided by locally constrained TD algorithms. A simple linear initialization is proposed for event functions and a refinement algorithm for target vectors that advances one event at a time. Locally constrained TD methods incur higher spectral distortion that may be contained by increasing the linear prediction (LP) analysis frame rate used to obtain the LSF tracks.**

***Index Terms*— line spectral frequency, LSF, LSP, linear prediction, speech coding, temporal decomposition.**

I. Introduction

Speech coders for commercial communication services traditionally use uniform sampling of speech parameter tracks. However, for low-bit-rate codecs uniform sampling is incompatible with toll quality speech reconstruction since speech transitions will not get enough definition. Therefore, it is important to track speech events and somehow allocate the coding bits between the time-domain evolution of parameter tracks and their frequency-domain resolution.

The idea of speech events may be implemented in different ways. Temporal decomposition (TD) is based on the notion that parameter tracks could be constructed with localized interpolation functions. It was proposed by Bishnu Atal to be used with LAR (log area ratio) tracks [1]. It has also been applied to LSF (line spectral frequency) tracks which are the most popular linear prediction (LP) representation in speech coding [2] and has been proposed for excitation parameters as well, such as pitch and voicing decisions [3].

II. Temporal decomposition based on singular value decomposition

Linear prediction analysis produces LSF vectors at a regular frame rate, describing p parameter tracks, where p is the order of LP analysis. Each set of consecutive frames for joint TD analysis will be referred to as a superframe. The first proposal uses 16 frame superframes. Then, for the locally constrained TD proposal, the length of the superframe is allowed to vary according to the rate of event detection.

The LSF evolution matrix $\mathbf{Y}$ contains pth order LSF vectors $\mathbf{y}_n$ for $n = 0, 1, \ldots N-1$ as its columns. It is temporally decomposed, generating target matrix $\mathbf{A}$ and event matrix $\mathbf{\Phi}$, which may be used to estimate $\mathbf{Y}$ as

$$\hat{\mathbf{Y}} = \mathbf{A}\mathbf{\Phi}, \tag{1}$$

The columns $\mathbf{a}_j$ in matrix $\mathbf{A}$ for $j = 0, 1, \ldots, J-1$ are the target vectors, where J is the number of events. Event functions $\phi_j(n)$ for $n = 0, 1, \ldots, N-1$ and $j = 0, 1, \ldots, J-1$ are represented as vectors $\boldsymbol{\phi}_j$ whose transposes are the rows in matrix $\mathbf{\Phi}$.

Bishnu Atal's original proposal for event function [1] selection is based upon a negative-slope zero-crossing criterion over the linear weighted displacement function

$$\nu(l) = \frac{\sum_{n=0}^{N-1}(n-l)\phi^2(n)}{\sum_{n=0}^{N-1}\phi^2(n)}. \tag{2}$$

The interpolated LSF vectors obtained after TD by applying Eq. (1) differ from the original ones by the interpolation error vectors, arranged as columns in matrix

$$\mathbf{E} = \mathbf{Y} - \mathbf{A}\mathbf{\Phi} \tag{3}$$

thus fitting the original LSF vectors $\mathbf{Y}$ within a total square error

$$\varepsilon_E = \text{tr}\left(\mathbf{E}^T\mathbf{E}\right) \tag{4}$$

By setting to zero the partial derivatives of ε_E with respect to the elements in target matrix $\mathbf{A}$, it will be reestimated as

$$\mathbf{A} = \mathbf{Y}\mathbf{\Phi}^T\left(\mathbf{\Phi}\mathbf{\Phi}^T\right)^{-1}. \tag{5}$$

Each reestimated target vector $\mathbf{a}_j$ has to be tested for stability by using the ordering property of LSFs, checking whether $a_{k,j} < a_{k+1,j}$ for $k = 1, 2, ..., p-1$. Should it happen that $a_{k,j} \geq a_{k+1,j}$, then they will be replaced by

$$\tilde{a}_{k,j} = \left(a_{k,j} + a_{k+1,j} - \varepsilon\right)/2 \tag{6}$$

$$\tilde{a}_{k+1,j} = \left(a_{k,j} + a_{k+1,j} + \varepsilon\right)/2 \tag{7}$$

This work is supported by Fundaç˜ao de Amparo à Pesquisa do Estado de S˜ao Paulo (FAPESP) under Grants no. 04/13206-7 and 05/52912-7.

where the LSF separation introduced [4] is $\varepsilon = 0.01$.

Additionally, given the reestimated target matrix, the partial derivatives of the square error in Eq. (4) with respect to the elements in event matrix $\mathbf{\Phi}$ are now set to zero, while retaining previous samples from the other event functions for the calculation, in order to obtain the reestimated event matrix elements as

$$\phi_j(n) = \frac{\sum_{k=1}^{p} y_k(n) a_{k,j} - \sum_{\substack{l=0 \\ l \neq j}}^{J-1} \phi_l(n) \sum_{k=1}^{p} a_{k,l} a_{k,j}}{\sum_{k=1}^{p} a_{k,j}^2} \quad (8)$$

for $n = 0, 1, \ldots, N-1$ and $j = 0, 1, \ldots, J-1$ so that, as suggested by Atal [1], in each refinement iteration Eq. (5) and Eq. (8) are computed in this order.

As the number of refinement iterations is increased, the log spectral distortion (SD) was found to initially decrease and then to start increasing. Therefore, it has been found advantageous to use this turning point as a criterion for a variable number of refinement iterations. Otherwise, refinement stops after a maximum of 10 iterations. It should be noted that this improves upon the original suggestion of four iterations [1]. Lower log SDs have been found by using this refinement procedure. For instance, in one case, an initial log SD of 12.30 dB dropped to 1.54 dB. More results are presented in Section IV.

As a further test of the distortion introduced by TD, the proposed method was applied to the LSF tracks in the MELP reference speech coder [5]. The quality of the reconstructed speech was measured by means of the PESQ (Perceptual Evaluation of Speech Quality) [6]. The additional degradation caused by TD was found to be as low as 0.061 MOS PESQ units.

However, a major drawback of the refinement is that event functions tend to spread wider around their centers, making them less attractive for speech coding. In order to counteract that tendency, a constrained refinement procedure is proposed below.

III. Locally constrained temporal decomposition

For the initial detection of event locations, the original TD minimizes the linear displacement measure in Eq. (2), which is based on event functions by means of the singular value decomposition of the original LSF evolution matrix $\mathbf{Y}$. A simpler measure, based on linear regression within a local window centered around each original LSF vector $\mathbf{y}(n)$, is the spectral feature transition rate (SFTR) [3], whose unnormalized version, the spectral transition measure (STM) [4], is

$$D_T(n) = \left\| \sum_{m=-M}^{M} m\mathbf{y}(n-m) \right\|^2 \quad (9)$$

with $M = 2$. The local minima of $D_T(n)$ are the internal event centers $C(j)$ for $j = 1, 2, \ldots, J-2$. Additionally, endpoint event centers are located at $C(0) = 1$ and $C(J-1) = N-1$. The detected event rate is

$$f_e = \frac{J}{N} f_f, \quad (10)$$

where f_f is the LP analysis frame rate. For high frame rates, the STM detects event centers at an adequate rate. But for low frame rates, the number of events detected is insufficient and the interpolation error is used to detect more events as presented further below.

The initial target vectors are identified to the LSF vectors at the event center locations, that is,

$$\mathbf{a}_j = \mathbf{y}(C_j) \quad (11)$$

for $j = 0, 1, \ldots, N-1$.

This locally constrained TD admits only two overlapping nonzero event functions that add up to unity at every sample and each one of them ranges from unity down to zero. Given the center locations of the event functions, the initial event functions are assumed to be straight lines between adjacent centers. For instance,

$$\phi_j(n) = \begin{cases} 1 - \frac{n - C(j)}{C(j+1) - C(j)} & \text{for } C(j) \leq n < C(j+1) \\ 1 + \frac{n - C(j)}{C(j) - C(j-1)} & \text{for } C(j-1) \leq n < C(j). \end{cases} \quad (12)$$

The local support of the event functions reduces the LSF vector estimates to lie within the superframe between the current event function center and the next one, that is, for superframe j, the LSF vector estimates are

$$\hat{\mathbf{y}}(n) = \mathbf{a}_j \phi_j(n) + \mathbf{a}_{j+1} \phi_{j+1}(n) \quad (13)$$

for $n = C(j), C(j)+1, \ldots, C(j+1)-1$ so that the interpolation error vectors within superframe j are

$$\begin{aligned} \mathbf{e}(n) &= \mathbf{y}(n) - \hat{\mathbf{y}}(n) \\ &= \mathbf{y}(n) - \mathbf{a}_j \phi_j(n) - \mathbf{a}_{j+1} \phi_{j+1}(n) \end{aligned} \quad (14)$$

and the total square interpolation error taken along superframe j is

$$\varepsilon_j = \sum_{n=C(j)}^{C(j+1)-1} \|\mathbf{e}(n)\|^2. \quad (15)$$

For refinement, only the right-hand target vector $\mathbf{a}_{j+1}$ is allowed to vary within a segment since it is assumed that the left-hand target vector $\mathbf{a}_j$ was reestimated in the refinement for the previous superframe. Therefore, the interpolation error vectors in Eq. (14) may be rewritten as

$$\mathbf{e}(n) = \mathbf{v}(n) - \mathbf{a}_{j+1} \phi_{j+1}(n), \quad (16)$$

where vectors

$$\mathbf{v}(n) = \mathbf{y}(n) - \mathbf{a}_j \phi_j(n) \quad (17)$$

for $n = C(j), C(j)+1, \ldots, C(j+1)-1$ are constant so that the total square error in Eq. (15) may be expanded as

$$\begin{aligned} \varepsilon_j &= \sum_{n=C(j)}^{C(j+1)-1} \|\mathbf{v}(n) - \mathbf{a}_{j+1} \phi_{j+1}(n)\|^2 \\ &= \sum_{n=C(j)}^{C(j+1)-1} \Big(\|\mathbf{v}(n)\|^2 - 2\mathbf{v}^T(n) \mathbf{a}_{j+1} \phi_{j+1}(n) \\ &\quad + \|\mathbf{a}_{j+1}\|^2 \phi_{j+1}^2(n) \Big). \end{aligned} \quad (18)$$

Taking the gradient of the total square error in Eq. (18) with respect to the right-hand target vector and setting it equal to the zero vector, the reestimated right-hand target vector is obtained as

$$\mathbf{a}_{j+1} = \frac{\sum_{n=C(j)}^{C(j+1)-1} \phi_{j+1}(n)\mathbf{v}(n)}{\sum_{n=C(j)}^{C(j+1)-1} \phi_{j+1}^2(n)} \tag{19}$$

Finally, substituting $\mathbf{v}(n)$ from Eq. (17) into Eq (19) and using vector notation, it results in the reestimate

$$\mathbf{a}_{j+1} = \frac{\left(\mathbf{Y}_j - \mathbf{a}_j\boldsymbol{\theta}_j^T\right)\boldsymbol{\phi}_{j+1}}{\left\|\boldsymbol{\phi}_{j+1}\right\|^2}, \tag{20}$$

where the columns in matrix $\mathbf{Y}_j$ are the LSF vectors $\mathbf{y}(n)$ and $\boldsymbol{\theta}$ is a local truncation of event vector $\boldsymbol{\phi}_j$ such that $\theta(n - C(j)) = \phi_j(n)$, in both cases for $n = C(j), C(j) + 1, \ldots, C(j+1) - 1$ in superframe j.

Once the reestimate of the right-hand target vector is computed according to Eq. (20), in order to complete a refinement iteration, the event functions lying in the superframe must be reestimated. By using the constraint that the two event functions add up to unity at every sample in the superframe, Eq. (14) may be rewritten as

$$\mathbf{y}(n) - \mathbf{a}_{j+1} = (\mathbf{a}_j - \mathbf{a}_{j+1})\,\phi_j(n) + \mathbf{e}(n), \tag{21}$$

thus reducing the problem to the reestimation of the right-hand side of the left-hand event function. Demanding orthogonality between the difference vector $\mathbf{a}_j - \mathbf{a}_{j+1}$ and the interpolation error vector $\mathbf{e}(n)$, the raw estimate $\hat{\boldsymbol{\phi}}_j$ is obtained as

$$\hat{\phi}_j(n) = \frac{(\mathbf{y}(n) - \mathbf{a}_{j+1}))^T (\mathbf{a}_j - \mathbf{a}_{j+1})}{\|\mathbf{a}_j - \mathbf{a}_{j+1}\|^2} \tag{22}$$

for $n = C(j), C(j)+1, \ldots, C(j+1)-1$. Since the values obtained by using Eq. (22) may fall outside the zero to unity range [4], [2], the estimate in Eq. (22) is modified to

$$\phi_j(n) = \min\left\{1, \max\left\{0, \hat{\phi}_j(n)\right\}\right\} \tag{23}$$

for $n = C(j), C(j)+1, \ldots, C(j+1)-1$.

We have found that the event rate obtained by searching for the local minima of the STM is high enough when the LSF vectors result from LP analysis with largely overlapping frames. But, when the frame overlapping is low, the event rate obtained with STM alone is too low and, then, the local maxima of total square error in Eq. (15) are also declared as event centers [4]. Examples of simple and improved event localization are presented in Section IV.

IV. Experiments with temporal decomposition

A study on the number of necessary event functions for LSF track TD was carried out for the SVD-based TD. The frame rate for LP analysis was set to $f_f = 44.44$ Hz without superposition between consecutive frames. The signals from dialect region DR1 in the test partition of the TIMIT database were used resampled at 8 kHz, including 70 utterances from male speakers and 40 utterances from female speakers for a total of 351.2 s of speech. The set of LSF vectors in a superframe was projected onto lower dimension subspaces of dimensions 10 down to one. Average, minimum and maximum event rates are shown in Table I along with the average log SD for each initial subspace dimension and the average, maximum and minimum statistics for the number of refinement iterations.

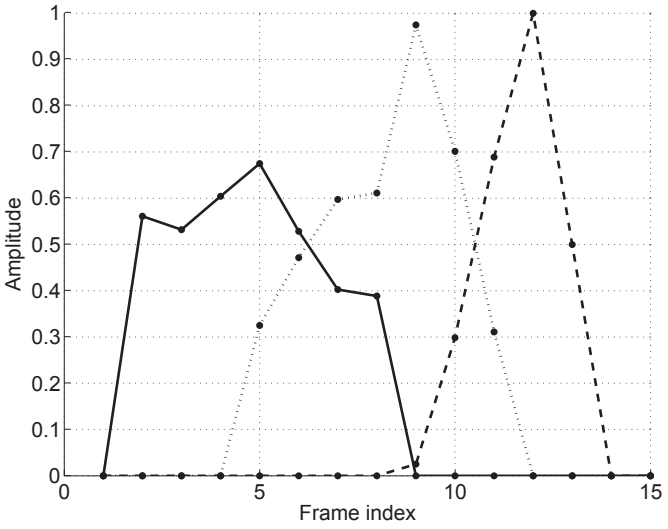

Fig. 1. Three consecutive event functions obtained by locally constrained TD with improved event localization for an LSF track.

As can be seen in Table I, the greater the singular subspace dimension and the number of event functions the lower the spectral distortion.

The locally constrained TD algorithm presented in Section III using the simple event localization procedure was tested for LP analysis frame rate $f_f = 500$ Hz and frame length $L = 200$ samples.

As test signals for the locally constrained methods, all the 1680 sentences in the test partition of the TIMIT speech database have been used resampled at 8 kHz, including 1120 utterances from male speakers and 560 utterances from female speakers for a total of 5186.7 s of speech. The log spectral distortion (SD) measure has been used for comparing the spectra for TD interpolated LSFs with those represented by the original LSFs.

The event rate distribution for locally constrained TD with simple event localization is outlined in Table II by its average, maximum and minimum values while the interpolation error distribution is sketched by the fraction of outliers above log SD levels ranging from 2 dB to 8 dB in 2 dB steps. Refinement for the locally constrained methods has been carried out for up to two iterations.

The locally constrained TD algorithm presented in Section III using the improved event localization procedure was tested under two LP analysis conditions. Their frame rates are $f_f = 44.44$ Hz and $f_f = 500$ Hz and both employ frame length $L = 200$ samples. Three consecutive event functions, obtained at the lower frame rate, are illustrated in Fig. 1. The lower frame rate and the frame length are compatible

TABLE I

PERFORMANCE OF TEMPORAL DECOMPOSITION BASED ON SVD FOR SEVERAL LSF SINGULAR SUBSPACE DIMENSIONS.

Singular subspace dimension	Average event rate (s^{-1})	Max. event rate (s^{-1})	Min. event rate (s^{-1})	Log SD (dB)	Number of iterations		
					Max.	Min.	Average
10	12.09	19.44	5.56	1.48	10	2	7.58
9	11.18	19.44	2.78	1.63	10	2	7.50
8	10.28	16.67	2.78	1.79	10	2	7.37
7	9.46	16.67	2.78	1.95	10	2	7.08
6	8.70	16.67	2.78	2.11	10	2	6.96
5	7.91	13.89	2.78	2.30	10	1	6.61
4	7.10	11.11	2.78	2.54	10	1	6.26
3	6.00	8.33	2.78	2.92	10	1	4.88
2	4.49	5.56	2.78	3.62	10	1	2.92
1	2.78	2.78	2.78	4.74	10	1	1.68

TABLE II

PERFORMANCE OF LOCALLY CONSTRAINED TD WITH SIMPLE EVENT LOCALIZATION FOR LOW AND HIGH LP FRAME OVERLAPPING.

Frame rate (s^{-1})	Average event rate (s^{-1})	Max. event rate (s^{-1})	Min. event rate (s^{-1})	Log SD (dB)	> 2 dB (%)	> 4 dB (%)	> 6 dB (%)	> 8 dB (%)
500.00	26.01	250.00	1.00	2.60	53.37	18.62	5.95	1.75

TABLE III

PERFORMANCE OF LOCALLY CONSTRAINED TD WITH IMPROVED EVENT LOCALIZATION FOR LOW AND HIGH LP FRAME OVERLAPPING.

Frame rate (s^{-1})	Average event rate (s^{-1})	Max. event rate (s^{-1})	Min. event rate (s^{-1})	Log SD (dB)	> 2 dB (%)	> 4 dB (%)	> 6 dB (%)	> 8 dB (%)
44.44	11.22	22.22	1.78	3.60	73.73	35.47	14.16	5.18
500.00	73.83	250.00	4.35	1.16	15.69	1.63	0.27	0.07

with the standard mixed excitation linear prediction (MELP) vocoder [5].

Before performing comparisons between spectral distortions caused by interpolation errors, it is important to keep in mind that interpolation error is almost never considered as an isolated source of degradation when comparing speech coders. Either the target spectral vectors are compared or the overall speech quality is measured between the original and the reconstructed speech as in our example at the end of Section II.

The SVD-based method provides superior performance. When using just 5 singular vectors, its distortion, as shown in Table I, is almost matched by the simple locally constrained method in Table II because it uses a higher frame rate. The simple method has the advantages of lower computational complexity and streaming operation. However, its event rate is higher, which may be a drawback for coding, even though it contributes to a more uniform performance.

Lower distortions can be attained when the improved localization criterion is used as shown in Table III for the higher frame rate. However, the event rate gets extremely high. By reducing the frame rate, a more manageable event rate can be obtained, but at a much higher distortion. Nonetheless, it should be noted that a high event rate may be interesting as a tool for spectral analysis and may also be useful for coding if coupled to a good sampling tool.

V. CONCLUSION

Temporal decomposition algorithms for LSF parameter tracks have been implemented and analyzed, using the original SVD method as a reference, whose behavior has been tested for different numbers of singular vectors since this number is directly related to the event rate and quality obtained. Also, the number of singular vectors used sets an upper bound to the event rate which can be effectively obtained. It has been observed that the number of refinement iterations may be lower or higher than that suggested originally, motivating the proposal of a criterion to identify the best number. Aiming at a higher control of the extent and shape of event functions, locally constrained TD methods and criteria have been explored. A simple linear initialization of event functions is proposed as well as a sequential refinement of target vectors that advances the decomposition one event function at a time. Spectral distortion is higher for the local methods but it can be decreased by increasing the LP frame rate.

REFERENCES

[1] B. S. Atal, "Effi cient coding of LPC parameters by temporal decomposition," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, vol. 1, Boston, 1983, pp. 81–84.

[2] P. C. Nguyen and M. Akagi, "Improvement of the restricted temporal decomposition method for line spectral frequency parameters," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, vol. 1, Orlando, 2002, pp. 265–268.

[3] A. C. R. Nandasena and M. Akagi, "Spectral stability based event localizing temporal decomposition," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, vol. 2, Seattle, 1998, pp. 957–960.

[4] S.-J. Kim and Y.-H. Oh, "Effi cient quantisation method for LSF parameters based on restricted temporal decomposition," *IEE Electronics Letters*, vol. 35, no. 12, pp. 962–964, June 1999.

[5] L. M. Supplee, R. P. Cohn, J. S. Collura, and A. V. McCree, "MELP: The new Federal Standard at 2400 bps," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, vol. 2, Munich, 1997, pp. 1591–1594.

[6] ITU-T, "*Perceptual evaluation of speech quality (PESQ), an objective method for end-to-end speech quality assessment of narrowband telephone networks and speech codecs*," Recommendation P.862, Geneva, Feb. 2001.