Proc. ICASSP, pp. 5066-5069, 2010.

NON-NEGATIVE TEMPORAL DECOMPOSITION OF SPEECH PARAMETERS

Sadao Hiroya

NTT Communication Science Laboratories, NTT Corporation
3-1 Morinosato-Wakamiya, Atsugi-shi, Kanagawa, 243-0198 Japan

ABSTRACT

We present a non-negative temporal decomposition method for line spectrum pair and articulatory parameters. Based on the multiplicative update rules derived from a non-negative matrix factorization algorithm, these parameters decompose into a set of temporally overlapped event functions that are restricted to the range $[0, 1]$ and corresponding event vectors. With the proposed method, the RMS error of the measured and estimated articulatory parameters is 0.16 mm and the spectral distance of the measured and estimated line spectrum pair parameters is 1.97 dB. These results also show that these estimation errors of proposed method are significantly smaller than those of the conventional method. This technique will be useful for many applications, such as speech coding and speech modification.

Index Terms— Non-negative matrix factorization, Temporal decomposition, articulatory parameters

1. INTRODUCTION

In speech processing, it is important to efficiently represent multidimensional time-varying speech parameters, such as line spectrum pair (LSP) and articulatory parameters. A hidden Markov model (HMM) [1] is one of the models that can represent dynamical behavior as well as the trajectory smoothness of speech parameters. Temporal decomposition (TD) [2] can represent speech parameters as a set of temporally overlapped event functions and corresponding event vectors. TD has been used for many applications: speech coding [2, 3], segmentation of speech signals [4, 5], analysis of articulatory parameters [6, 7] and modification for the speech spectrum [8] as well as for the speaking rhythm [9]. For these purposes, the event functions should be restricted to the range $[0, 1]$. This restriction contributes to reducing the quantization error in speech coding and modifying speech parameters easily. However, they have been clipped at the range $[0, 1]$ because it is difficult to apply the restriction to the conventional TD algorithm directly. Thus, a novel TD algorithm with the restriction is required.

In this paper, we present a non-negative temporal decomposition (NTD) method to overcome the problem. For given speech parameters, this method can optimize the event functions, which are restricted to the range $[0, 1]$, and the event vectors using the multiplicative update rules derived from a non-negative matrix factorization (NMF) [10] algorithm.

2. CONVENTIONAL TEMPORAL DECOMPOSITION

The original temporal decomposition [2] approximates the ith speech parameter $y_i(t)$ of time t to

$$\hat{y}_i(t) = \sum_{k=1}^{m} a_{i,k}\phi_k(t), \quad 1 \le t \le T,\ 1 \le i \le p \tag{1}$$

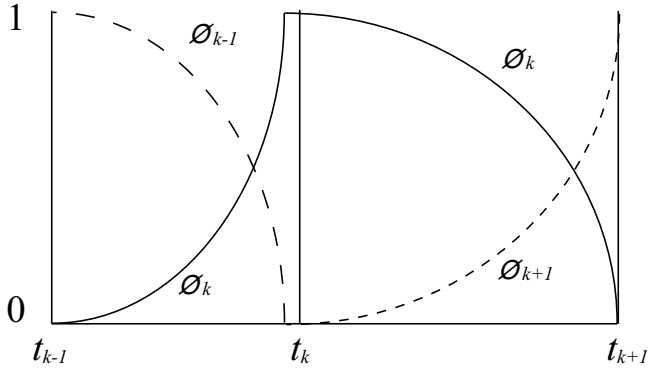

Fig. 1. Example of event functions.

where $a_{i,k}$ is the kth event vector, $\phi_k(t)$ is the kth event function, p is the dimension of the speech parameter, T is the length of the parameter sequence and m is the number of event function. Shiraki [11] has assumed that $\phi_k(t)$ is zero for $t < t_{k-1}$ and $t > t_{k+1}$ (see Fig. 1). Thus, $\hat{y}_i(t)$ can be represented as a linear combination of $a_{i,k}$ and $a_{i,k-1}$:

$$\hat{y}_i(t) = a_{i,k}\phi_k(t) + a_{i,k-1}\phi_{k-1}(t), \quad t_{k-1} \le t \le t_k. \tag{2}$$

Kim [3] has proposed that the sum of all event functions is one at any time t:

$$\hat{y}_i(t) = a_{i,k}\phi_k(t) + a_{i,k-1}(1 - \phi_k(t)). \tag{3}$$

The event functions $\phi_k(t)$ are determined by minimizing the least-squares error between $y_i(t)$ and $\hat{y}_i(t)$:

$$\phi_k(t) = \frac{\sum_{i=1}^{p}(a_{i,k} - a_{i,k-1})(y_i(t) - a_{i,k-1})}{\sum_{i=1}^{p}(a_{i,k} - a_{i,k-1})^2}. \tag{4}$$

However, the obtained event functions are not restricted to the non-negative value, nor the range $[0, 1]$. Thus, Kim clipped event functions at the range $[0, 1]$ [i.e. $\phi_k(t) = \min(1, \max(0, \phi_k(t)))$] and then updated event vectors $a_{i,k}$ using the least-squares method as follows.

$$\sum_{k=1}^{m} a_{i,k} \sum_{t=1}^{T} \phi_k(t)\phi_j(t) = \sum_{t=1}^{T} y_i(t)\phi_j(t), \quad 1 \le j \le m \tag{5}$$

$$\phi_k(t) \leftarrow \frac{\sum_{i=1}^{p} a_{i,k} y_i(t) + \alpha}{\sum_{i=1}^{p} (a_{i,k-1} a_{i,k} \phi_{k-1}(t) + a_{i,k}^2 \phi_k(t)) + \alpha(\phi_{k-1}(t) + \phi_k(t))} \phi_k(t) \tag{7}$$

$$\phi_{k-1}(t) \leftarrow \frac{\sum_{i=1}^{p} a_{i,k-1} y_i(t) + \alpha}{\sum_{i=1}^{p} (a_{i,k-1} a_{i,k} \phi_k(t) + a_{i,k-1}^2 \phi_{k-1}(t)) + \alpha(\phi_{k-1}(t) + \phi_k(t))} \phi_{k-1}(t) \tag{8}$$

$$a_{i,k} \leftarrow \frac{\sum_{t=t_{k-1}}^{t_{k+1}} y_i(t) \phi_k(t)}{\sum_{t=t_{k-1}}^{t_k - 1} (a_{i,k-1} \phi_{k-1}(t) \phi_k(t) + a_{i,k} \phi_k^2(t)) + \sum_{t=t_k}^{t_{k+1}} (a_{i,k+1} \phi_k(t) \phi_{k+1}(t) + a_{i,k} \phi_k^2(t))} a_{i,k} \tag{9}$$

3. NON-NEGATIVE TEMPORAL DECOMPOSITION

It is difficult to determine the event functions that are restricted to the range $[0, 1]$ by the least-squares method without a clipping, but the non-negative matrix factorization (NMF) [10, 12] algorithm can iteratively determine the non-negative value of event functions. Moreover, if we assume the model in Eq. (2) for $y_i(t)$ and $\phi_k(t) + \phi_{k-1}(t) = 1$ at any time t, we can determine the event functions that are restricted to the range $[0, 1]$. The problem is defined as

$$\begin{aligned} &\min \sum_{k=2}^{m} \sum_{t=t_{k-1}}^{t_k} \sum_{i=1}^{p} (y_i(t) - a_{i,k}\phi_k(t) - a_{i,k-1}\phi_{k-1}(t))^2 \\ &\text{s.t. } \phi_k(t) \geq 0,\ \phi_{k-1}(t) \geq 0,\ a_{i,k} \geq 0,\ a_{i,k-1} \geq 0 \\ &\quad \phi_k(t) + \phi_{k-1}(t) = 1 \quad \text{for } \forall t, i, k \end{aligned}$$

Thus, we consider minimizing the following cost function by NMF algorithm:

$$\begin{aligned} &\sum_{k=2}^{m} \sum_{t=t_{k-1}}^{t_k} \sum_{i=1}^{p} (y_i(t) - a_{i,k}\phi_k(t) - a_{i,k-1}\phi_{k-1}(t))^2 \\ &\qquad + \alpha \sum_{k=2}^{m} \sum_{t=t_{k-1}}^{t_k} (\phi_k(t) + \phi_{k-1}(t) - 1)^2, \end{aligned} \tag{6}$$

where $1 = t_1 < t_2 < \cdots < t_m = T$ and α is the weight. In line with a previous idea [12], we can obtain the multiplicative update rule for the event function and vector [Eqs. (7)-(9)].

4. OPTIMIZATION OF EVENT TIMING

The distortion $d(y(t), \hat{y}(t))$ of the cost function for each interval $t_{k-1} \leq t \leq t_k$ only depends on time t_{k-1} and t_k. Therefore, the event timing

$$t_k = \arg \min_{t_2, \cdots, t_{m-1}} \sum_{t=1}^{T} d(y(t), \hat{y}(t)), \quad 2 \leq k \leq m-1 \tag{10}$$

that minimizes total distortion for the whole interval $1 \leq t \leq T$ is derived efficiently by utilizing the dynamic programming (DP) method [11]. That is, we have

$$D(t_k) = \min_{t_{k-1} \in R_{k-1}} \Big(D(t_{k-1}) + \sum_{t=t_{k-1}}^{t_k} d(y(t), \hat{y}(t)) \Big) \tag{11}$$

where

$$R_{k-1} = \{t | t_{k-1} - \delta \leq t \leq t_{k-1} + \delta\}, \tag{12}$$

$D(t_k)$ is an accumulated minimal distortion at t_k and δ is a search range.

5. ALGORITHMS

5.1. Restricted TD (RTD)

In RTD [3], we first set the initial event vectors as $a_{i,k} = y_i(t_k)$ for the initial event timing t_k. Then, the event functions and the event timings are determined using Eq. (4) and DP. Finally, the event functions are clipped at the range $[0, 1]$ and only the event vectors are updated by Eq. (5) for given event functions and event timings. Note that Kim [3] has proposed a method for holding an ordering property of LSP parameters. But this is irrelevant to the main subject, thus we don't use this in the paper.

5.2. NTD

In NTD, the initial event vectors and event timings are the same as RTD. Then, the event functions and the event timings are determined by DP and the NMF update rules of Eqs. (7)-(8). Note that the event vectors are not updated at this stage. Finally, for given event timings, the event functions and vectors are updated using Eqs. (7)-(9).

6. EXPERIMENTAL CONDITIONS

We compare the proposed method (NTD) with the conventional method (RTD). We evaluate the proposed method using LSP and articulatory parameters. Articulatory parameters and speech signal

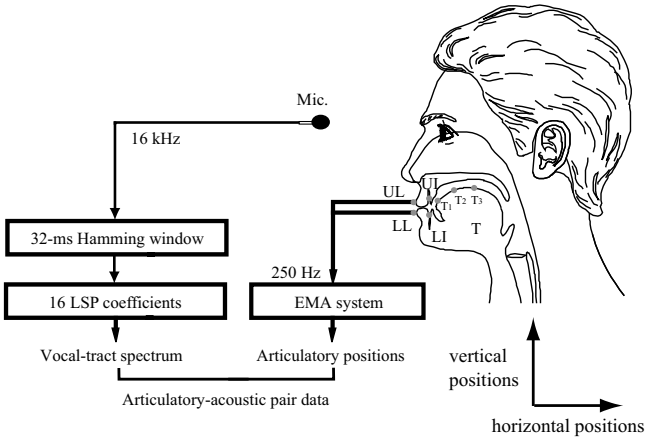

Fig. 2. Simultaneous observations to obtain articulatory-acoustic data.

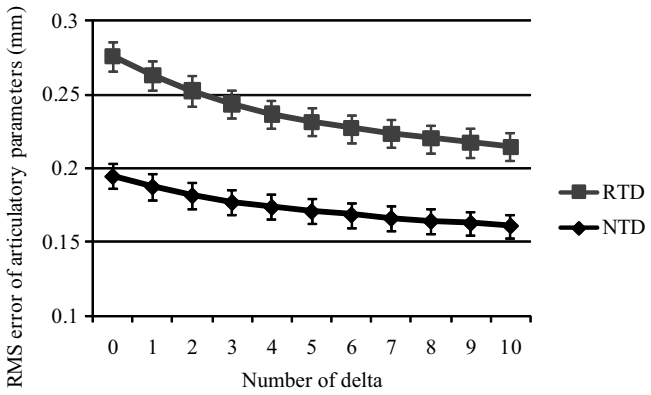

Fig. 3. RMS error of articulatory parameters for the number of δ's. Vertical bars indicate the standard error of the mean.

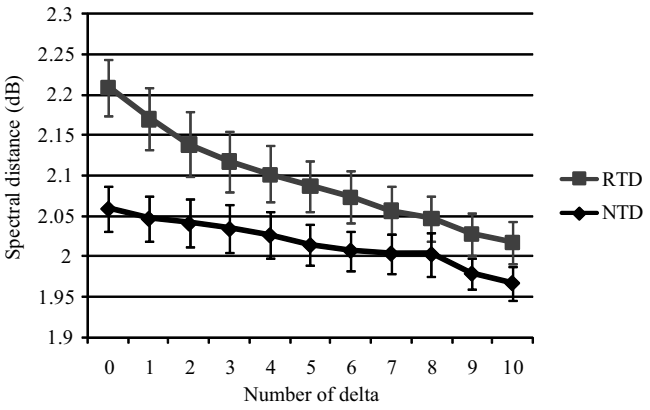

Fig. 4. Spectral distance of LSP parameters for the number of δ's. Vertical bars indicate the standard error of the mean.

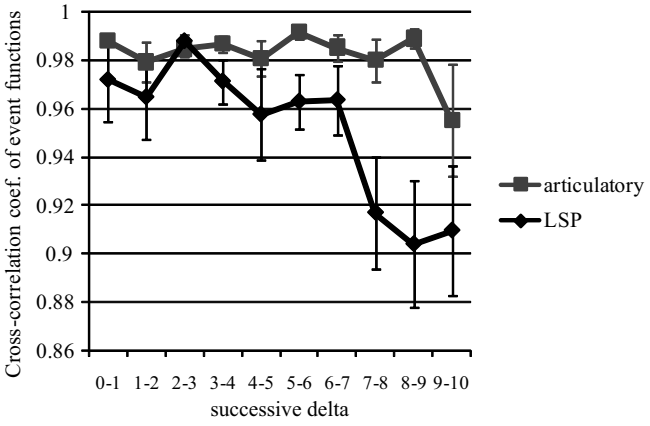

Fig. 5. Cross-correlation coefficients of event functions. Vertical bars indicate the standard error of the mean.

data were obtained from simultaneous observations using an electromagnetic articulographic (EMA) system [1] and audio recording of continuous speech utterances (Fig. 2). The articulatory data were collected using the EMA at a sampling rate of 250 Hz. The articulatory parameters were represented by the vertical and horizontal positions of six receiver coils, which were placed on the lower incisor (LI), the upper and lower lips (UL, LL), and the tongue (T1, T2, T3; three positions). The speech signal was recorded at a sampling rate of 16 kHz. Sixteen LSP parameter coefficients without the 0-th coefficient were obtained using a 32-ms Hamming window with a 4-ms frame shift.

In articulatory-acoustic recordings, one Japanese male subject read 16 Japanese sentences. The average numbers of phonemes and frames per sentence were 37 and 748, respectively. We labeled the central point of each phoneme on the time axis, which we call initial event timing t_k. Thus, m is the number of phonemes. The labeling was done manually by an expert. The α was set to 10^6 and 10^3 for articulatory and LSP parameters, respectively.

7. RESULTS

Figure 3 shows the estimation error of articulatory parameters using RTD and NTD for the number of δ's. For every δ, the error of NTD was smaller than that of RTD. Figure 4 shows the spectral distance of LSP parameters for the number of δ's. For every δ, the spectral distance of NTD was smaller than that of RTD. For both articulatory and LSP parameters, the estimation error using NTD monotonically decreased for the number of δ's. For NTD, $\phi_k(t) + \phi_{k-1}(t)$ at any time t was one. This indicates that the proposed method efficiently reduces the estimation error under the condition that the event functions are restricted to the range $[0, 1]$.

The estimation error of articulatory parameters for NTD (0.16 mm) was much smaller than 1.22 mm for an articulatory HMM [13] and 1.65 mm for a kinematic triphone model [14] because NTD parameters have not been statistically modeled in this study. But we expect that the error of NTD with statistical modeling will be equivalent to or smaller than that of other articulatory modeling by using redundancy in NTD parameters.

For the application of NTD to speech coding and speech modification, better stability of the event function would contribute to reducing the quantization error and improving the quality of modified speech. To investigate this, we assessed the cross-correlation coefficients of event functions for the successive δ's for articulatory and LSP parameters (Fig. 5). This was done by concatenating the event functions as $[\phi_1, \phi_3, \phi_5, \cdots]$: we calculated the cross-correlation of solid lines in Fig. 6(d) and (e). The coefficient of the articulatory

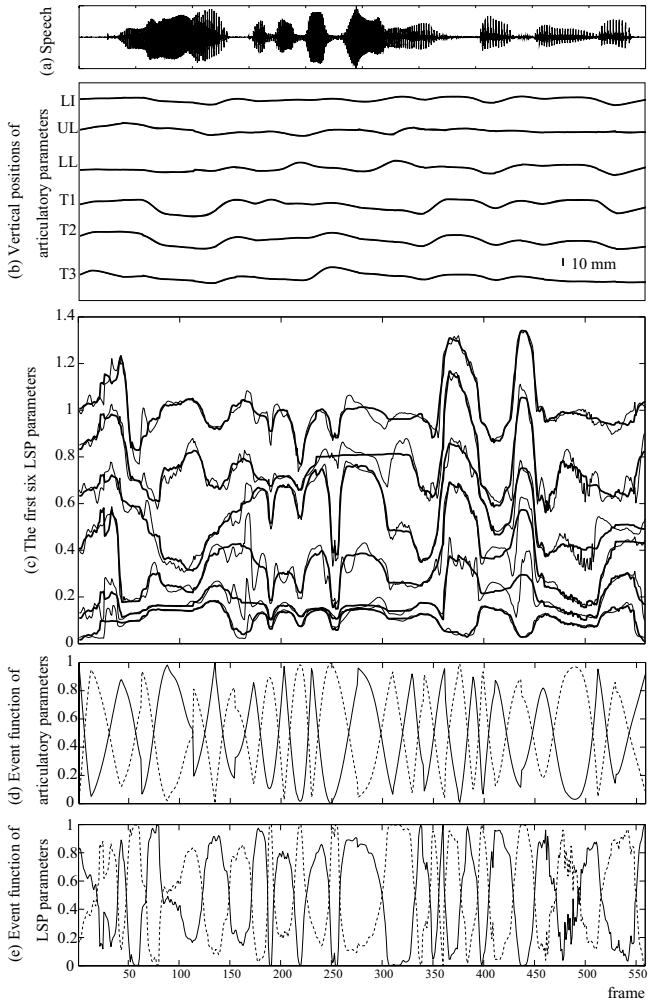

Fig. 6. (a) Speech waveform. (b) Measured (thin lines) and estimated (thick lines) articulatory parameters of vertical positions. (c) Measured (thin lines) and estimated (thick lines) first six LSP parameters. (d) Event functions of articulatory parameters. (e) Event functions of LSP parameters. $\delta = 5$.

parameters is almost 0.99, but that of LSP is smaller. Moreover, we found the event functions of LSP parameters were largely changed for δ of more than 7. This indicated that the stability of event functions of articulatory parameters using NTD is better than that of LSP parameters. One possible reason is that the temporal patterns of the articulatory parameters are simpler and smoother than those of the LSP parameters, but a more detailed analysis is required.

Finally, the cross-correlation coefficient of event functions between LSP (Fig. 6(e)) and articulatory parameters (Fig. 6(d)) using NTD was calculated. The coefficients were 0.65, 0.53, and 0.36 for $\delta = 0, 5$, and 10, respectively. The lag was around 3 msec.

8. CONCLUSIONS

The paper presented a non-negative temporal decomposition method for speech parameters. The error of NTD is smaller than that of RTD under the condition that the event functions are restricted to the range $[0, 1]$. Using NTD, we plan to analyze speaker variability of articulatory parameters.

9. ACKNOWLEDGEMENTS

The author thanks Drs. H. Gomi and K. Kameoka for many useful and helpful discussions.

10. REFERENCES

[1] Hiroya, S. and Honda, M., "Estimation of Articulatory Movements from Speech Acoustics Using an HMM-Based Speech Production Model," *IEEE Trans. Speech Audio Process.*, pp. 175–185, 2004.

[2] Atal, B.S., "Efficient Coding of LPC Parameters by Temporal Decomposition," in *ICASSP*, 1983, pp. 81–84.

[3] Kim, S-J. and Oh, Y-H., "Efficient Quantisation Method for LSF Parameters Based on Restricted Temporal Decomposition," *Electronics Letters*, pp. 962–964, 1999.

[4] Bimbot, F., Chollet, G., Deleglise, P., and Montacie, C., "Temporal Decomposition and Acoustic-phonetic Decoding of Speech," in *ICASSP*, 1988, pp. 445–448.

[5] Dix, P.J. and Bloothooft, G., "A Breakpoint Analysis Procedure Based on Temporal Decomposition," *IEEE Trans. Speech Audio Process.*, pp. 9–17, 1994.

[6] Jung, T-P., Krishnamurthy, A.K., Ahalt, S.C., Beckman, M.E., and Lee, S-H., "Deriving Gestural Scores from Articulatory-Movement Records Using Weighted Temporal Decomposition," *IEEE Trans. Speech Audio Process.*, pp. 2–18, 1996.

[7] Collins, M.J., Krishnamurthy, A.K., and Ahalt. S.C., "Generating Gestural Scores from Articulatory Data Using Temporal Decomposition," *IEEE Trans. Speech Audio Process.*, pp. 230–233, 1999.

[8] Nguyen, B.H. and Akagi, M., "A Flexible Spectral Modification Method Based on Temporal Decomposition and Gaussian Mixture Model," *Acoust. Sci. & Tech.*, pp. 170–179, 2009.

[9] Shiraki, Y., "Optimal Temporal Decomposition for Voice Morphing Preserving Δ Cepstrum," *IEICE Trans. Fundamentals*, pp. 577–583, 2004.

[10] Lee, D.D. and Seung, H.S., "Learning the Parts of Objects by Non-negative Matrix Factorization," *Nature*, pp. 788–791, 1999.

[11] Shiraki, Y. and Honda, M., "Extraction of Temporal Pattern of Spectral Sequence Based on Minimum Distortion Criterion," in *ASJ Spring Meeting*, 1991, pp. 233–234.

[12] Virtanen, T., "Monaural Sound Source Separation by Nonnegative Matrix Factorization With Temporal Continuity and Sparseness Criteria," *IEEE Trans. Audio Speech and Lang. Process.*, pp. 1066–1074, 2007.

[13] Hiroya, S. and Mochida, T., "Multi-speaker Articulatory Trajectory Formation Based on Speaker-independent Articulatory HMMs," *Speech Commun.*, pp. 1677–1690, 2006.

[14] Okadome, T. and Honda, M., "Generation of Articulatory Movements by Using a Kinematic Triphone Model," *J. Acoust. Soc. Am.*, pp. 453–463, 2001.