ON MODELING EVENT FUNCTIONS IN TEMPORAL DECOMPOSITION BASED SPEECH CODING

S. Ghaemmaghami *M. Deriche* *B. Boashash*

Signal Processing and Avionic Research Centre
Queensland University of Technology
2 George st, Brisbane, Q 4001, Australia
shahrokh@markov.eese.qut.edu.au m.deriche@qut.edu.au b.boashash@qut.edu.au

ABSTRACT

Temporal Decomposition (TD) is an efficient technique for modeling speech spectral evolution through orthogonalization of the matrix of spectral parameters which reduces the amount of spectral information in TD-based speech coding. We have shown in earlier work that "event" functions can be approximated by fixed-width Gaussian functions with a minor degradation in the reconstructed speech, leading to further bit-rate reduction in such systems. In this paper, through perceptually-based spectral distortion measurement, we show the impact of events shape on the speech quality, and propose a new composite function and discuss its effect on the coder performance using different combinations of spectral parameters in event detection and speech synthesis.

1. INTRODUCTION

1.1. Temporal decomposition

Temporal Decomposition (TD) is a method to model the phonemic evolution of speech on the basis of a time sequence of spectral parameters, in the *least square* sense [1]. The phonemic evolution is presented by a number of time-overlapping compact functions, called *target* or *event* functions, which are interpreted as physical representations of speech *acoustic events* [1]. TD maps time trajectories of spectral parameters into event space through a set of linear equations:

$$\hat{y}_i(n) = \sum_{k=1}^{m} a_{ik}\phi_k(n), \ 1 \leq n \leq N, \ 1 \leq i \leq p \qquad (1)$$

where p is the number of parameters of each frame, n is the frame index, $\hat{y}_i(n)$ is the ith parameter approximated, $\phi_k(n)$ is the kth event at frame n, a_{ik} is the weighting factor, and m the number of events in the interval $n = 1$ to $n = N$.

1.2. Approximation of Event functions

In previous work [2], we approximated the event functions using fixed-width (σ) Gaussian functions, by accepting a minor degradation in speech quality. Using such an approximation, we only needed the event locations to place the Gaussian functions. This led to a considerable reduction in the number of bits required for coding the events. Moreover, the event refinement, which is a time-consuming task, can be eliminated and events are approximated as:

$$\hat{\phi}_k(n) = exp(-(n-n_c)^2/2\sigma^2), \qquad 1 \leq n \leq N \quad (2)$$

which are non-zero only in the interval assigned to the relevant segment (the segment in which n_c is the index for the central frame).

Although Gaussian function yields a good approximation to the shape of most *original* events [2], it had some drawbacks. The major one being its limited flexibility in shape as it is exclusively characterized by σ, such that any change in *wideness* at the centroid affects the effective width of the event and vice versa.

In this paper, on the basis of our previous findings, we develop a composite non-linear function to get a better approximation to event functions. We will shown here, based on performance evaluation through perceptually-based spectral distortion measurement of the reconstructed speech, that a considerable improvement in speech quality and a lesser sensitivity to the parameters of the approximating function is achieved using this composite function.

2. THE PROPOSED METHOD

There are three basic features to be considered for approximating event functions: *compactness*, *curvature*,

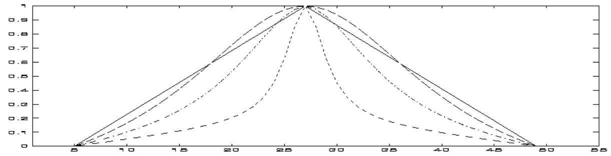

Figure 1. The proposed function with different values for *curving* and *sharpening* parameters.

and *sharpness*, which are specified by *width*, *curving*, and *sharpening* parameters, respectively. Each of these parameters has a specific role in reconstruction of spectral parameters. The *width* parameter designates the number of frames contributing to the corresponding event, the *curving* parameter denotes the *weighting* of spectral parameters at frames around the event centroid, and *sharpening* parameter mainly changes the amount of concentration on the event centroid.

To analyse these features and examine their effects on the system performance, we first generate a non-linear function $U_1(n)$ as:

$$U_1(n) = \begin{cases} \frac{(1+m)a}{(n-n_c)^2+a} - m, & |n-n_c| < \frac{N-1}{2} \\ 0, & elsewhere \end{cases} \quad (3)$$

$$m = \frac{4a}{(N-1)^2}$$

where a is the *curving* parameter, N is total number of frames within the segment, n is the frame index, and n_c is the index of central frame in the segment.

The curving parameter, a, changes the curvature of function U_1, mostly at its centroid.

The second function we construct, is a triangle function U_2 of the same length as U_1:

$$U_2(n) = \begin{cases} 1 - \frac{1}{\frac{(N-1)}{2}}|n-n_c|, & |n-n_c| < \frac{N-1}{2} \\ 0, & elsewhere \end{cases} \quad (4)$$

As evident, U_1 and U_2 are symmetric functions equal to zero at two ends and 1 at the center (maximum). Now we form the composite function U as follows:

$$U = (1-b)U_1 + bU_2 \quad (5)$$

where b is the *sharpening* parameter, between zero and 1, such that larger b gives a sharper function.

This function can form functions with different wideness or sharpness, as Gaussian (with zero ends) and triangle shapes as well, by choosing appropriate values for a and b (see figure 1).

3. DISTORTION MEASURE

The system performance was assessed through a perceptually based spectral distance measure using *modified Bark-scaled filter-bank* [3], as:

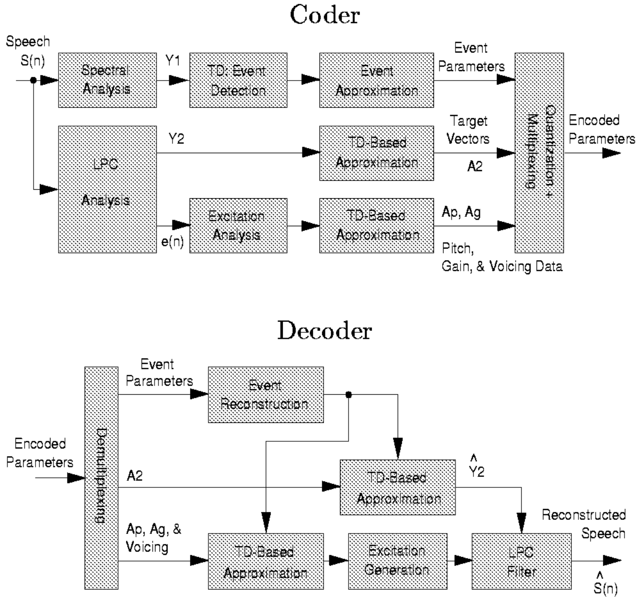

Figure 2. Block diagram of the coding system developed.

$$d_s = \frac{1}{N\,S_1^{av}} \sum_{i=1}^{N} \sum_{k=1}^{15} |S_1^i(k) - S_2^i(k)| \quad (6)$$

where i is the frame index, N is the total number of frames, S_1^{av} is the average power of the original speech over the whole utterance, $S_1^i(k)$ and $S_2^i(k)$ are the powers of original and encoded speech signals for the ith frame at the kth filter of the Bark-scale filterbank, whose center frequencies are given as:

$$f_0 = 600 sinh \frac{\nu + 1/2}{6}, \qquad \nu = 1, 2, ..., 15. \quad (7)$$

4. THE CODING SYSTEM

The structure of the coding system is shown in figure 2. The event functions are extracted from parameters matrix $\mathbf{Y}_1$ which can be *Cepstrum*, LA (*Log Area*), or LAR (*Log Area Ratio*) obtained from the "spectral analysis" block. The parameters set, $\mathbf{Y}_2$, is used in computing the matrix of *target vectors*, $\mathbf{A}_2$. The set $\mathbf{Y}_2$, referred to as speech synthesis set, is to be reconstructed at the decoder which will be used as LPC reconstruction filter coefficients for speech synthesis.

The excitation signal is approximated using *binary* model in an LPC system. The parameters of excitation signal, *pitch*, *gain*, and *voicing*, are compressed using the method proposed in [7]. In this method, pitch and gain contours are approximated through interpolation using the values at event centroids. This reduces the amount of excitation information to less than 1/10 with an acceptable accuracy, specifically for pitch (see [7]).

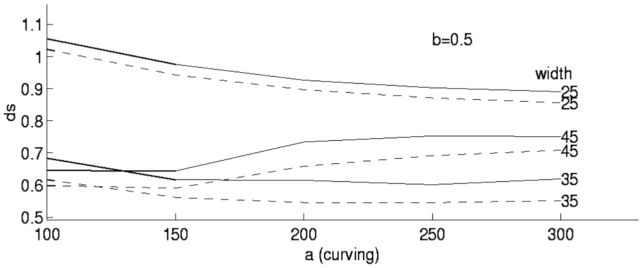

Figure 3. Spectral distance versus *curving* parameter, a, for *composite* function with different *widths*. Solid: Cep-LA, dashed: Cep-LAR.

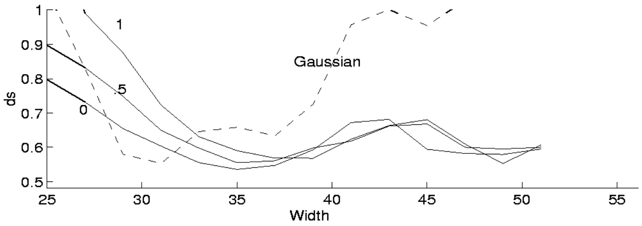

Figure 4. Spectral distance versus *width*, for *composite* function with different *sharpening* factor, b (solid curves). The same results are shown for Gaussian (dashed, abscissa: 4σ) for comparison.

At the decoder, the matrix of spectral parameters, $\mathbf{Y}_2$, is approximated as described in sections 1 and 2, and used as filter coefficients in an LPC model excited by the excitation signal reconstructed through the abovementioned TD-based interpolation to generate the reconstructed speech.

The system works at a rate around 600 b/sec as described in [2] where the proposed function used as event approximating function instead of Gaussian and the same values for frame-length, segment-length and overlap percentage were used.

5. EXPERIMENTAL RESULTS

To evaluate the usefulness of the function proposed above and to analyse the effect of its three basic parameters on the system performance, we devised the following method. For each possible value of the sharpening factor, b, we calculate the spectral distance between the original and the reconstructed speech signals for different combinations of the other two parameters, a (curving) and N (width). The experiments were carried out using different spectral parameters in event detection and speech synthesis.

Some results are displayed in Figures 2-4 and Table 1, where all durational parameters, as N (width) and σ (for Gaussian function), are shown in terms of *frame-period* (T) which is equal to 5 msec, and total segment length in TD is taken equal to $55T$.

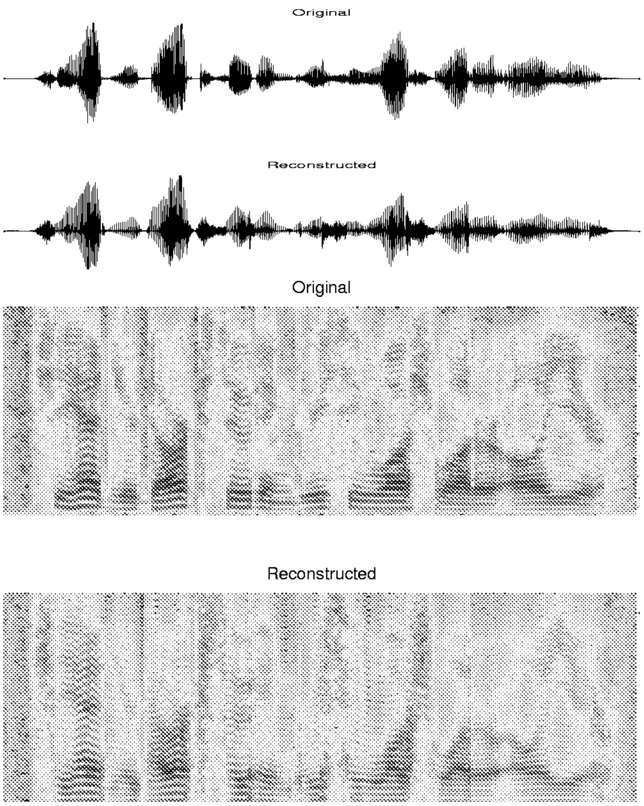

Figure 5. Waveform and spectrogram of original speech and speech reconstructed using the proposed function. The utterance is */she had your dark suit in greasy wash water all year/.*

Figure 3 shows the overall effect of curving parameter, a, on the performance, evaluated by the spectral distance measure, d_s, for $b = .5$ and different widths (printed beside the curves), where Cepstrum coefficients and LA (or LAR) parameters are used as detection and synthesis sets, respectively.

Figure 4 represents the spectral distance versus *width* parameters, N, for different sharpening factors, b, printed on the solid curves. The same result using Gaussian (dashed) function is also depicted for comparison.

The reconstructed speech waveform of an utterance and its spectrogram using the proposed function, with $curving = 200$, $sharpening = 0$, and $width = 35$, can be seen in figure 5 where *Cepstrum* and *LAR* parameters are used in event detection and speech synthesis, respectively.

To get an overall assessment of the proposed method, we examined a number speech samples, from TIMIT database, using the composite function. Table 1 shows a summary of results with the best values for different parameters and corresponding spectral distances obtained using equation (6).

Event→	Gaussian		Triangle		Composite			
Parameters↓	σ	d_s	w	d_s	N	b	a	d_s
Cep-LA	8	.60	35	.61	35	0	200	.58
Cep-LAR	8	.57	35	.56	35	0	200	.53
LA-LA	11	.77	47	.88	51	0	400	.83
LA-LAR	11	.69	47	.82	51	0	400	.73
LAR-LA	7	.66	41	.60	45	0	250	.60
LAR-LAR	7	.58	41	.57	45	0	250	.54

Table 1. Best values for parameters of different approximating events and the corresponding spectral distances.

6. DISCUSSION

The effect of *curving* parameter, a, on the spectral distortion for different widths of approximating functions is shown in figure 3. As seen, this parameter compensates the effects of width, to some extent, by widening or narrowing the peak of the function. The system is more sensitive to changes in a when the function is very narrow (width=25) or very wide (width=45) but for width=35, where the *effective* width is around 90 ms (close to the average phoneme length [4]), a lesser sensitivity is obtained.

The *sharpening* parameter, b, mostly affects the peak, indeed the sharpness, of the function, hence, controls the amount of concentration on the event centroid. The larger the sharpness is, the lesser the contribution of spectral parameters of frames around the centroid to the corresponding event is obtained. Accordingly, a sharper function would be more suitable to approximate shorter events and vice versa. This effect should, however, be considered in combination with the effect of the *curving* parameter, as curving and sharpening acts, basically, in a conflicting manner in this sense.

The *width* parameter specifies the number of frames contributing to the corresponding event centroid. This means that in order to emphasis on *phonetic invariant* instants [5], we require narrower approximating events. As original events are found with different widths, there is only a limited range of acceptable values for *width* parameter to keep spectral distortion in reconstructed speech below a predefined threshold. As illustrated in figure 4, the proposed function yields a lesser sensitivity to event width than that obtained with Gaussian function, which leads to better performance in speech and speaker independent coding.

As shown in figure 4, the best performance is obtained with $a = 200$, $b = 0$, and $width = 35$. This conforms with our informal subjective assessment in a 600 b/s TD-based coder implemented. However, we found that sharper functions ($b > 0$) could be more useful in approximating some short events, such as events related to *stop* consonants.

7. CONCLUSION

A composite function, to model speech events in TD-based speech coding, has been proposed in this paper. Through perceptually based spectral distance measurement, we have shown that with appropriate shaping parameters, better coding performance, or higher quality in reconstructed speech, can be achieved with the function compared to that obtained using Gaussian approximating events. It has also been shown that among different spectral parameters, the pair *Cepstrum-LAR*, as *detection* and *synthesis* sets, yields the lowest spectral distortion.

8. REFERENCES

[1] B.S. Atal, "Efficient Coding of LPC Parameters by Temporal Decomposition", *Proc. ICASSP 83*, pp. 81-84, 1983.

[2] S. Ghaemmaghami, M. Deriche, "A New Approach to Very Low-Rate Speech Coding Using Temporal Decomposition", *Proc. ICASSP'96*, Vol. 1, pp. 224-227, May 1996.

[3] S. Wang, A. Sekey, A. Gersho, "An Objective Measure for Predicting Subjective Quality of Speech Coders", *IEEE J. Select. Areas in Comm.*, Vol. 10, No. 5, pp. 819-829, June 1992.

[4] L.C.W. Pols, X. Wang, L.F.M. ten Bosch, "Modelling of Phone Duration (using the TIMIT database) and its Potential Benefit for ASR", *Speech Comm. 19*, pp. 161-176, 1996.

[5] S. Blumstein, K. Stevens, "Acoustic Invariance in Speech Production: Evidence from Measurements of the Spectral Characteristics of Stop Consonants", *J. Acoust. soc. Am.*, 66(4), pp. 1001-1017, Oct. 1979.

[6] Y.M. Cheng, D. O'Shaughnessy, "On 450-600 b/s Natural Sounding Speech Coding", *IEEE, Trans. SAAP*, Vol. 1, No. 2, pp. 207-219, 1993.

[7] S. Ghaemmaghami, M. Deriche, "A New Approach to Approximation of Pitch and Gain Contours Using Temporal Decomposition", *Proc. 6th Australian Regional AES Convention (AES'96)*, Sept. 1996.