EFFICIENT COMPRESSION OF MELP SPECTRAL PARAMETERS USING OPTIMIZED TEMPORAL DECOMPOSITION

Chandranath N. Athaudage, Alan B. Bradley and Margaret Lech

School of Electrical and Computer Systems Engineering
Royal Melbourne Institute of Technology, Australia

Abstract

This paper describes an efficient compression algorithm for MELP (Supplee *et al.*, 1997) spectral parameters based upon an optimized temporal decomposition model of speech. Temporal decomposition (TD) is an effective technique for modelling the dynamics of speech parameters and an optimized algorithm for TD has been presented previously (Athaudage, Bradley & Lech, 1999). In this paper we provide an overview of the optimized TD algorithm with its rate-distortion performance. Application of optimized TD for efficient compression of MELP spectral parameters is discussed with TD parameter quantization issues and effective coupling between TD analysis and parameter quantization stages. Simulation results show that over 50% compression can be achieved using 450 ms delay block coding, which is attractive for speech storage related applications.

1 INTRODUCTION

Temporal decomposition (TD) of speech (Atal 1983; Ghaemmaghami & Deriche, 1996; Nandasena & Akagi, 1998; Athaudage, Bradley & Lech, 1999) is a technique of modelling speech parameter trajectories in terms of a sequence of acoustic event targets and an associated sequence of event functions. TD can also be considered as an effective method of decorrelating the inherent inter-frame correlation present in any frame based parametric representation of speech. In previous work with TD (Athaudage, Bradley & Lech, 1999) we have proposed a dynamic programming based optimization strategy for a modified temporal decomposition model of speech with higher modelling accuracy. The technique is termed optimized TD because acoustic events within speech blocks are optimally located unlike in the previous TD methods (Atal 1983; Ghaemmaghami & Deriche, 1996; Nandasena & Akagi, 1998) where a-priori assumptions like spectral stability are made on event locations. Section 2 of this paper provides an overview of this optimized TD algorithm and its performance. A methodology of incorporating the optimized TD for efficient low bit rate coding of MELP spectral parameters at the expense of 450 ms coding delay is described in Section 3. Conclusions and further research targets are given in Section 4.

2 OPTIMIZED TD ALGORITHM

The modified TD model of speech (Athaudage, Bradley & Lech, 1999) considered for optimization allows only two event functions (two adjacent events) to overlap at any given frame location. This model can be expressed as given in Eq. (1).

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

where, $\mathbf{a}_k$, $\phi_k(n)$ and n_k are the kth event target vector, the amplitude the kth event function at the nth frame, and the center location of the kth event function, respectively. The vector $\hat{\mathbf{y}}(n)$ is the approximation of the nth spectral parameter vector $\mathbf{y}(n)$, produced by the TD model. The index n represents the discrete frame number. Equivalently, this TD model can be expressed as follows.

$$\hat{\mathbf{y}}(n) = \sum_{k=1}^{K} \mathbf{a}_k \phi_k(n), \quad 1 \leq n \leq N \tag{2}$$

$$\text{where,} \quad \phi_k(n) = 0 \text{ for } n < n_{k-1} \text{ and } n \geq n_{k+1}$$

The total set of speech parameters, $\mathbf{y}(n)$, where $1 \leq n \leq N$, buffered for TD analysis is termed a *block*. N and K are the number of frames per block and the number of events per block, respectively. The analysis is done on *block-by-block* basis, and for each block the event locations, event targets and event functions are optimally evaluated. For optimal performance a buffering technique with overlapping blocks is used to ensure a smooth transition of events at the block boundaries. N and K determine the level of TD resolution and can be expressed in terms of events per second as K/N times the frame rate. Given a block of speech parameters, $\mathbf{y}(n)$, where $1 \leq n \leq N$, and the number of events, K, to be located within the block, first, the event location set, $\{n_1, n_2, \ldots, n_K\}$, and the event function set $\{\vec{\phi}_1, \vec{\phi}_2, \ldots, \vec{\phi}_K\}$, are jointly evaluated using a dynamic programming based optimization strategy. Initially, $\mathbf{a}_k = \mathbf{y}(n_k)$ is used as an approximation for event targets. Then the event target set, $\{\mathbf{a}_1, \mathbf{a}_2, \ldots, \mathbf{a}_K\}$, is refined by re-evaluating them to minimize the reconstruction error of the TD model. The final reconstruction error for the nth frame, D_n, in terms of spectral distortion (dB) can be evaluated as given in Eq. (3).

$$D_n = \sqrt{\frac{1}{2\pi} \int_{-\pi}^{\pi} [10\log_{10}(\mathrm{S}(\mathrm{e}^{\mathrm{j}\omega})) - 10\log_{10}(\hat{\mathrm{S}}(\mathrm{e}^{\mathrm{j}\omega}))]^2 \mathrm{d}\omega} \tag{3}$$

Where, $S(e^{j\omega})$ and $\hat{S}(e^{j\omega})$ are the frequency responses of the LPC filter corresponding to $\mathbf{y}(n)$ and $\hat{\mathbf{y}}(n)$, respectively. An important feature of the optimized TD algorithm is its ability to freely select an arbitrary number of events per block, i.e. average number of events per second (event rate). This was not the case in previous TD algorithms (Atal 1983; Ghaemmaghami & Deriche, 1996; Nandasena & Akagi, 1998), where the number of events was limited by other constraints like spectral stability. Average event rate, also called the TD resolution, determines the reconstruction error (distortion) of the TD model. Lower distortion levels can be expected for higher TD resolution. However, higher resolution implies a lower compression efficiency in application point of view. Therefore, the rate-distortion characteristic of the optimized TD algorithm is quite important for coding applications, and simulations were carried out to determine this. A speech data set consists of 16 phonetically-diverse sentences of TIMIT speech database was used as the test set. MELP spectral parameters, i.e. line spectral frequencies (LSF), calculated at 22.5 ms frame intervals were used as the speech parameters for TD analysis. Average spectral distortion was evaluated for the event rates of 4, 8, 12, 16, 20 and 24 events/sec. Figure 1 shows the average spectral distortion versus event rate graph. Base frame rate point, i.e. 44.4 frame/sec, is also shown for reference. The significance of the frame rate is that if the event rate is made equal to frame rate, therefore in this case 44.44 events/sec, theoretically the average speactral distortion should become zero. This is the maximum possible TD resolution and corresponds to situation where all the event functions become

unit impulses spaced at frame intervals, and the event targets become exactly equal to the original spectral parameter frames.

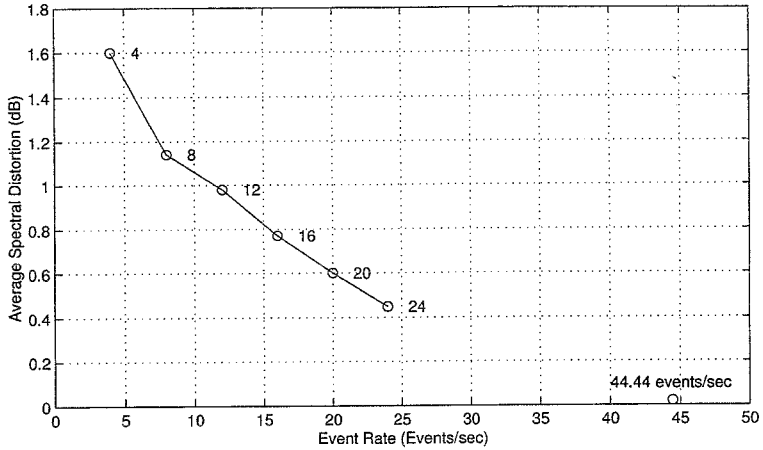

Figure 1. Average spectral distortion (dB) versus TD resolution (events/sec) characteristic of the optimized TD algorithm.

3 OPTIMIZED TD BASED MELP SPECTRAL PARAMETER CODING

3.1 MELP Coder

The Mixed Excitation Linear Prediction (MELP) (Supplee *et al.*, 1997) coder is the U.S. Department of Defence coding standard at 2.4 kbits/sec. It uses LPC filter to model the speech spectrum and a mixed excitation consisting of gain, pitch, Fourier magnitudes and bandpass voicing parameters to model the LPC residual. Both spectral and excitation parameters are calculated at 22.5 ms frame intervals. This corresponds to the base frame rate of 44.44 frames/sec. The spectral parameter set consists of the 10^{th} order LPC coefficients converted to Line Spectral Frequencies (LSFs). A 25-bit multi-stage vector quantizer (MSVQ) (Leblanc *et al.*, 1993) is used for the LSF quantization. A total of 54 bits/frame is required to code all the mixed excitation parameters and LSFs. Therefore, nearly half (25 out of 54) of the number of bits is used for the spectral parameter coding. Sections 3.2 through 3.5 describe the details of the proposed new coding scheme for the MELP spectral parameters based upon the optimized TD algorithm. This scheme requires an additional buffering delay of 450 ms and therfore is only meant for voice storage related applications.

3.2 Block Schematics

Figure 2 shows the proposed TD based coding scheme for MELP spectral parameters. This operates on 20-frame block (450 ms) mode and replaces the frame based MSVQ stage for spectral parameters of the standard MELP coder. The block size was set to 20 frames because investigations showed that further increase of block size gives deminishing returns

in terms of TD model accuracy while increasing the computational time in the order of N^2, where N is the block size. The TD resolution was set to 12 events/sec as a compromize between the rate and the distortion as described in Section 2. Coupling between the TD analysis and the parameter quantization stages is used to achieve better post-quantation parameter reconstruction accuracy. Refienment stage in Figure 2, which re-evaluates the event targets based on the quantized event functions, indicates this coupling.

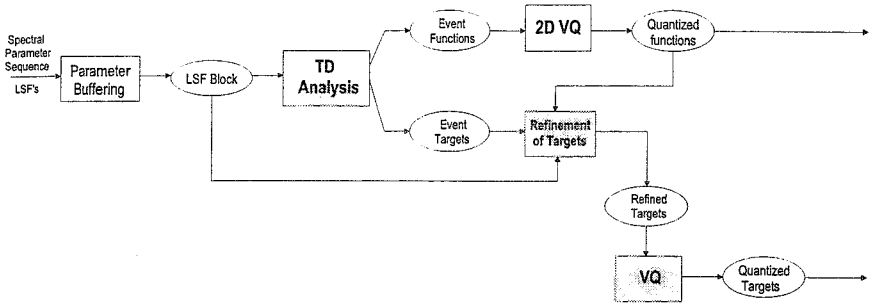

Figure 2. Proposed optimized temporal decomposition based MELP spectral parameter compression scheme. This is to replace the frame based MSVQ stage of the standard MELP coder.

3.3 Event Function Quantization

One choice for quantization of the event function set, $\{\vec{\phi}_1, \vec{\phi}_2, \ldots, \vec{\phi}_K\}$, for each block is to use vector quantization (VQ) (Gersho & Gray, 1992) on individual event functions, $\vec{\phi}_k$'s, in order to exploit any dependancies in event function shapes. However, the event functions are of variable legnth and therefore require normalization to a fixed length before vector quantization. Investigations showed that the process of normalization/denormalization itself introduces a considerable error which gets added to the quantization error. Therefore, we incorporated a frame based 2 dimensional VQ for event functions which proved to be simple and effective. This was possible only because the modified TD model allows only two event functions to overlap at any frame location. Vectors, $[\phi_k(n)\ \ \phi_{k+1}(n)]^T$, were quantized individually. The distribution of the 2 dimensional vector points of $[\phi_k(n)\ \ \phi_{k+1}(n)]^T$ showed significant clustering, and this dependancy was effectively exploited through the frame level vector quantization of event functions. Thirty two phonetically-diverse sentences from TIMIT database resulting in 4428 LSF frames were used as the training set to generate the code books of sizes 5, 6, 7, 8 and 9-bits, using the LBG k-means algorithm (Linde, Buzo & Gray, 1980).

3.4 Event Target Quantization

Quantization of the event target set, $\{\mathbf{a}_1, \mathbf{a}_2, \ldots, \mathbf{a}_K\}$, for each block was performed by vector quantizing each target vector, $\mathbf{a}_k$, separately. Event targets are 10 dimensional LSFs, but they differ from the original LSFs due to the iterative refinement of the event targets incorporated in the TD analysis stage. VQ code books of sizes 6, 7, 8 and 9-bits were generated using the same training data set described in Section 3.3.

3.5 Objective Performance Evaluation

Spectral parameters can be synthesized from the quantized event targets, $\hat{\mathbf{a}}_k$'s, and quantized event functions, $\hat{\phi}_k$'s, for each speech block as given in Eq. (4).

$$\hat{\mathbf{y}}(n) = \sum_{k=1}^{K} \hat{\mathbf{a}}_k \hat{\phi}_k(n), \quad 1 \leq n \leq N \tag{4}$$

Where, $\hat{\mathbf{y}}(n)$ is the nth synthesized spectral parameter vector at the decoder, synthesized using the quantized TD parameters. The average error between the original spectral parameters, $\mathbf{y}(n)$'s, and the synthesized spectral parameters, $\hat{\mathbf{y}}(n)$'s, calculated in terms of average spectral distortion (dB) was used to evaluate the objective quality of the coder. The final bit rate requirement for spectral parameters of the proposed compression scheme can be expressed in number of bits per frame, as given in Eq. (5).

$$B = n_1 + n_2 \frac{K}{N} + n_3 \frac{K}{N} \quad \text{bits/frame} \tag{5}$$

where, n_1 and n_2 are the sizes (in bits) of the code books for the event function quantization and event target quantization, respectively. The parameter n_3 denotes the number of bits required to code each event location within a given block. It was found that event location information, which is always an integer between 1 and 20, can be losslessly coded using differential encoding with $n_3 = 4$. As described in section 3.2 the block size, N, and the TD resolution were set to 20 frames/block and 12 events/sec, respectively. This makes the number of events per block, $K = 5$. The average spectral distortion (SD) was evaluated for different combinations of the sizes of event function and event target code books. The same test speech data set described in Section 2 was used. Figure 3 shows the average SD (dB) for different n_1 and n_2 against B (bits/frame). Note that n_3 is fixed at 4-bits.

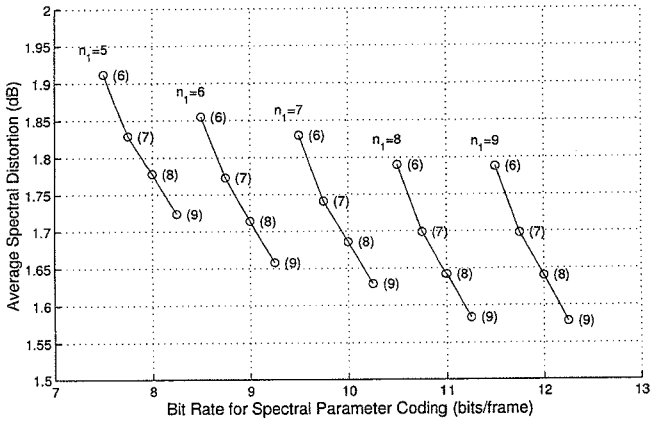

Figure 3. Average spectral distortion after TD parameter quantization. n_1 is the code book size for event function quantization. Code book size for event target quantization, n_2, is depicted as (n_2).

It is important to note that due to the nature of the quantization technique incorparated for the event functions and coupling between the TD analysis and the TD parameter quantization stages, the variation of the average spectral distortion with n_1 is made minimal. Simulation results using scalar quantization of the event functions confirmed the advantages of the proposed event function quantization technique. An average SD of about 1.62 dB is achievable for $n_1 = 7$, $n_2 = 9$ and $n_3 = 4$. This corresponds to the bit rate of $B = 10.25$ bits/frame, compared to the 25 bits/frame rate of the original MELP coder with 1.22 dB of average SD. This indicates over 50% compression at the expense of 0.4 dB of objective quality and 450 ms of processing delay. Informal listening tests showed that this small degradation in the objective quality is practically negligible.

4 CONCLUSION

An efficient compression scheme for MELP spectral parameters based upon optimized TD algorithm has been proposed. The TD parameter quantization technique and the coupling between the TD analysis and the parameter quantization stages can be highlighted as the main features of the compression scheme. Simulation results show that over 50% compression ratio can be achieved at the expense of 450 ms delay in processing, which is useful for voice storage ralated applications. The formal subjective performance evaluation of the coder remains for future research.

REFERENCES

Atal B.S. (1983), "Efficient coding of LPC parameters by temporal decomposition", ICASSP'83, Boston, pp. 81-84.

Athaudage C.N., Brabley A.B., & Lech M. (1999), "Optimization of a temporal decomposition model of speech", ISSPA'99, Brisbane, Australia, pp. 471-474.

Gersho A. & Gray R.M. (1992), *Vector quantization and signal compression*, Kluwer Academic Publishers, Dordrcht, Netherlands.

Ghaemmaghami S. & Deriche M. (1996), "A new approach to very low rate speech coding using temporal decomposition", ICASSP'96, pp. 224-227.

Leblanc W.P., Bhattacharya B., Mahmond S.A., & Cuperman V. (1993), "Efficient search and design procedures for robust multi-stage VQ of LPC parameters for 4 kb/s speech coding", IEEE Transactions on Speech and Audio Processing, vol. 1, No. 4, pp. 373-385.

Linde, Y., Buzo A., & Gray R.M. (1980), "An Algorithm for vector quantiser design", IEEE Transactions on Communications, vol. 28, pp. 84-95.

Nandasena A.C.R. & Akagi M. (1998), "Spectral stability based event localizing temporal decomposition", ICASSP'98, Seattle, USA, pp. 957-960.

Supplee L.M., Cohn R.P., Collura J.S., & McCree A.V. (1997), "MELP: The new federal standard at 2400 bps", ICASSP'97, pp. 1591-1594.