UNIWERSYTET TECHNOLOGICZNO-PRZYRODNICZY
IM. JANA I JĘDRZEJA ŚNIADECKICH
W BYDGOSZCZY

WYDZIAŁ TELEKOMUNIKACJI
I ELEKTROTECHNIKI

ZESZYTY NAUKOWE NR 250

# TELEKOMUNIKACJA I ELEKTRONIKA 11

BYDGOSZCZ – 2008

UNIWERSYTET TECHNOLOGICZNO-PRZYRODNICZY
IM. JANA I JĘDRZEJA ŚNIADECKICH
W BYDGOSZCZY

ZESZYTY NAUKOWE NR 250

# TELEKOMUNIKACJA I ELEKTRONIKA 11

BYDGOSZCZ – 2008

REDAKTOR NACZELNY
prof. dr hab. inż. Janusz Prusiński

REDAKTOR DZIALOWY
dr inż. Sławomir Cieślik

OPRACOWANIE TECHNICZNE
mgr inż. Daniel Morzyński

ISSN 1899-0088

Wydawnictwa Uczelniane Uniwersytetu Technologiczno-Przyrodniczego
ul. Ks. A. Kordeckiego 20, 85-225 Bydgoszcz, tel. (052) 3749482, 3749426
e-mail: wydawucz@utp.edu.pl http://www.utp.edu.pl/~wyd

Wyd. I. Nakład 80 egz. Ark. aut. 4,75. Ark. druk. 5,00.
Zakład Małej Poligrafii UTP Bydgoszcz, ul. Ks. A. Kordeckiego 20

# Contents

UNIWERSYTET TECHNOLOGICZNO-PRZYRODNICZY
IM. JANA I JĘDRZEJA ŚNIADECKICH W BYDGOSZCZY
ZESZYTY NAUKOWE NR 250
TELEKOMUNIKACJA I ELEKTRONIKA 11 (2008), 5-31

# EFFICIENT TRANSCODING OF AN MPEG-2 BIT STREAM TO AN H.264 BIT STREAM

Rochelle Pereira[1], K. R. Rao[2], Att Kruafak[2]

[1]NVIDIA Corporation
Santa Clara, CA 95050USA
rpereira81@yahoo.com

[2]Electrical Engineering Department
The University of Texas at Arlington
416 Yates Street, Box 19016
Arlington, Texas 76019, USA
{rao, att.kruafak}@uta.edu

*Summary*: The objective of this paper is to develop a technique for transcoding from MPEG-2 main profile to H.264 main profile and compare its performance with other transcoding architectures. The proposed transcoder reuses information from the MPEG-2 bit stream taking into account the improved techniques such as multiple block size motion estimation, in loop deblocking filter, intra directional prediction, integer DCT, context adaptive variable length coding, adaptive weighted prediction, and human visual sensitivity (HVS) weighting, adopted in H.264. The proposed method achieves low complexity, comparable quality and reduced bit rate in the transcoding process compared to previous techniques.

Keywords: MPEG-2, H.264, video transcoding

## 1. INTRODUCTION

MPEG-2 [1,2] has been a widely accepted video coding standard for various applications ranging from DVD to digital TV broadcast. A large variety of products based on the MPEG-2 standard are available in the market. The most important goal of MPEG-2 is to make the storage and transmission of digital audio visual material more efficient. The latest standard, H.264 AVC (advanced video coding) [3–5], has an even broader perspective to support high and low bit rate multimedia applications on existing and future networks. The advantage in terms of better quality at a lower bit-rate is why H.264 is fast replacing MPEG-2 [6]. However, the user end hardware such as set-top-boxes, DVD players had previously been adapted for MPEG-2 coded bit streams. This gives rise to a need for portability between MPEG-2 and H.264.

Video transcoding is the operation of converting video from one format to another [7–10]. A format is defined by characteristics such as bit-rate, spatial resolution, etc. One of the earliest applications of transcoding is to adapt the bit-rate of a compressed stream to the channel bandwidth for universal multimedia access in various channels like wireless networks, Internet, dial-up networks, etc. Changes in the characteristics of an encoded bit stream like bit-rate, spatial resolution, quality, etc., can also be achieved by scalable video coding [8]. However, in cases where the available network bandwidth is insufficient or if it fluctuates with time, it may be difficult to set the base layer bit-rate. In addition, scalable video coding demands additional complexities at both the encoder and the decoder. One such complexity is coding the bit stream in such a way that parts of the bit steam available can be extracted at the decoder based on the available bandwidth and the decoder processing power and display constraints.

Section 2 describes the various transcoding architectures that have been proposed in the literature, followed by their advantages and disadvantages addressed in section 3. The transcoding architecture proposed in this paper is described in section 4. Subsections 4.1, 4.2 and 4.3 devote to transcoding of I (intra) frame, P (predicted) frame and B (bidirectionally interpolated) frame respectively. Section 5 summarizes the tests of the complete transcoder that includes I, P, and B frames. Conclusions and further research are listed in sections 6 and 7 respectively.

## 2. TRANSCODING ARCHITECTURES

The basic architecture for converting an MPEG-2 elementary stream into an H.264 elementary stream arises from complete decoding of the MPEG stream and then re-encoding into an H.264 stream. However, this involves significant computational complexity [7]. Hence there also is a need to transcode at low complexity. Transcoding can in general be implemented in the spatial domain or in the transform domain or in a combination of the two domains. Several transcoding architectures have been proposed earlier.

### OPEN LOOP TRANSFORM DOMAIN TRANSCODING (Fig. 1) [8]

Open loop transcoders are computationally efficient. They operate in the DCT domain. However they are subject to drift error. Drift error occurs due to rounding, quantization loss and clipping functions. This can also accumulate due to the mismatch in the IDCT implementation accuracy in the encoder and the decoder.

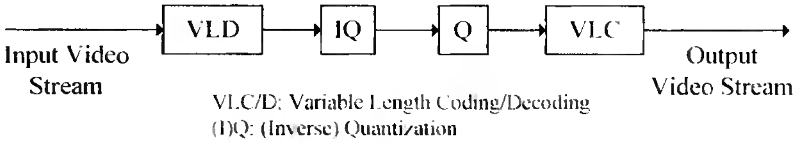

Input Video Stream
VLD
IQ
Q
VLC
Output Video Stream
VLC/D: Variable Length Coding/Decoding
(I)Q: (Inverse) Quantization

Fig. 1. Open loop transform domain transcoder architecture.

### CASCADED PIXEL DOMAIN ARCHITECTURE (CPDT) (Fig. 2) [8]

This is the most basic transcoding architecture. The motion vectors from the incoming bit stream are extracted and reused. Thus the complexity of the motion estimation block is eliminated which accounts for 60 % of the encoder computation [11]. As compared to the previous architecture, CPDT is drift free. Hence, even though it is slightly more complex, it is suitable for heterogeneous transcoding among different standards where the basic parameters like mode decisions, motion vectors, etc., are to be re-derived.

SIMPLIFIED DCT DOMAIN TRANSCODERS (SDDT) (Fig. 3) [8]

This transcoder is based on the assumption that DCT (discrete cosine transform), IDCT (inverse DCT) and motion compensation are linear processes. This architecture requires that motion compensation be performed in the DCT domain, which is a major computationally intensive operation [12]. For instance, as shown in Fig. 4, the goal is trying to compute the DCT coefficients of the target block B from the four overlapping blocks B1, B2, B3 and B4.

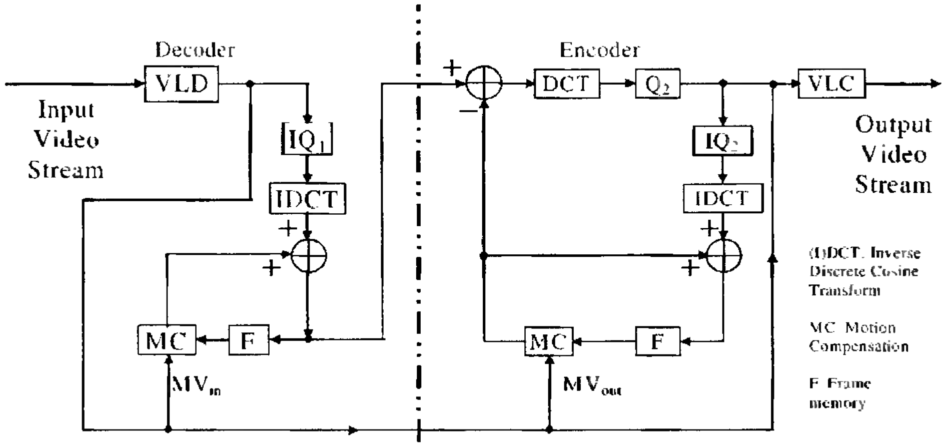


Fig. 2. Cascaded pixel domain transcoder architecture.

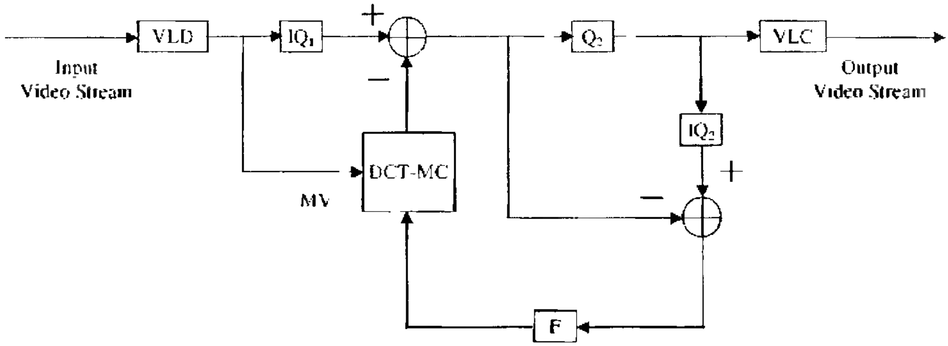


Fig. 3. Simplified transform domain transcoder architecture.

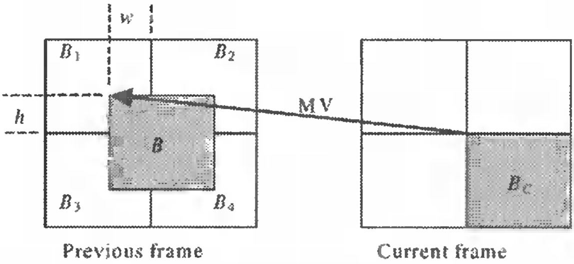


Fig. 4. Transform domain motion compensation illustration.

Also, clipping functions and rounding operations performed for interpolation in fractional pixel motion compensation lead to a drift in the transcoded video [13].

CASCADED DCT DOMAIN TRANSCODERS (CDDT) (Fig. 5) [8]

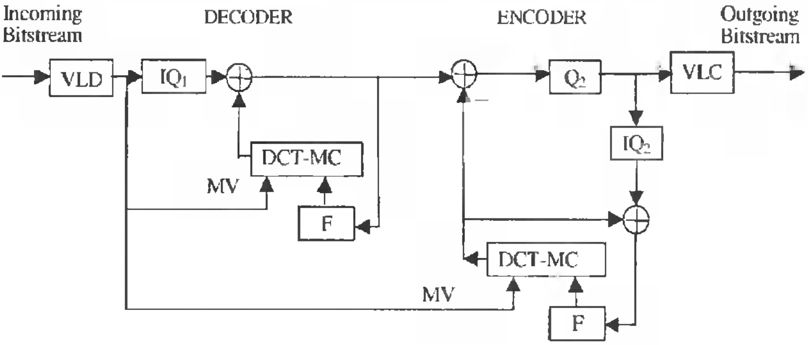


Fig. 5. Cascaded transform domain transcoder architecture.

This is used for spatial/temporal resolution downscaling and other coding parameter changes. As compared with SDDT, greater flexibility is achieved by introducing another transform domain motion compensation block; however it is far more computationally intensive and requires more memory [8]. It is often applied to downscaling applications where the encoder end memory will not cost much due to downscaled resolution.

## 3. CHOICE OF BASIC TRANSCODER ARCHITECTURE

DCT domain transcoders have the main drawback that motion compensation in transform domain is very computationally intensive. DCT domain transcoders are also less flexible as compared to pixel domain transcoders, for instance, the SDDT architecture can preferably be used for bit rate reduction transcoding. It assumes that the spatial and temporal resolutions stay the same and that the output video uses the same frame types, mode decisions and motion vectors as the input video.

For heterogeneous transcoding from MPEG-2 to H.264, it is required to implement several changes in order to accommodate the sophistication of H.264 as compared to

MPEG-2. For instance, MPEG-2 supports 16×16 and 16×8 macroblock partitions, but it is also required to refine the motion vectors to accommodate 8×16, 8×8 and sub 8×8 modes adopted in H.264. Hence, the use of DCT domain transcoders is not very practical.

From Figure 6, it can be inferred that the cascaded pixel domain architecture outperforms the DCT domain architecture. Also for larger GOP (group of pictures) (Fig. 7) sizes, the drift in DCT domain transcoders becomes more significant, as it progressively builds up till the next I frame is coded (Fig. 8). These large GOP sizes are practically used especially in implementations like networked video streaming and wireless video where high coding efficiency is desired.

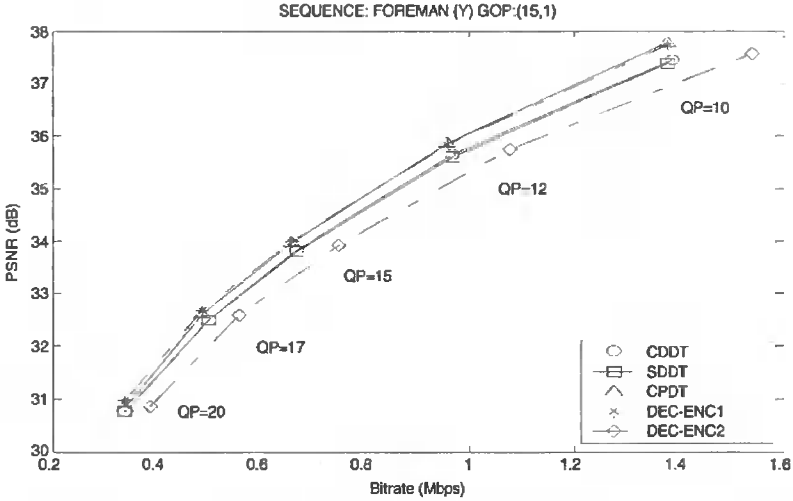


Fig. 6. PSNR vs. bit rate graph for the Foreman sequence encoded at QP = 7 and transcoded with different QP values and a GOP size 15, using different transcoding architectures as described in Figs. 2–4. DEC-ENC1 is CPDT using full scale full search motion estimation. DEC-ENC2 is CPDT using three step fast search motion estimation [8].

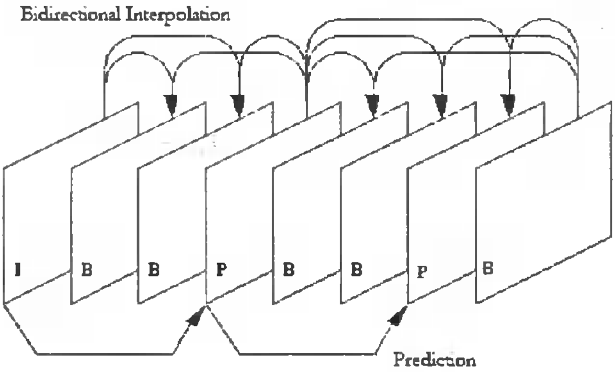


Fig. 7. A group of pictures, IBBPBBPB IBBPBBPB ···.

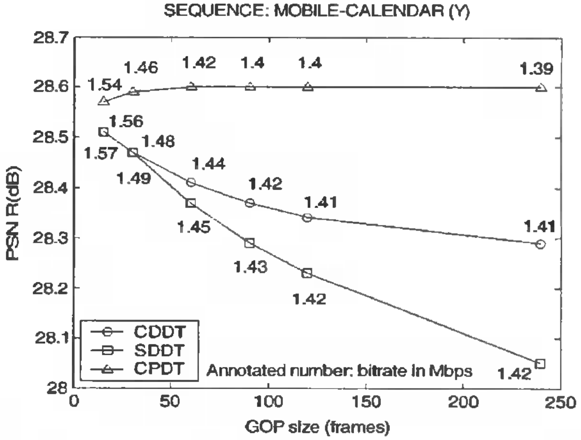


Fig. 8. Performance comparison of average PSNR for CPDT, SDDT and CDDT for different GOP sizes, using the test clip mobile-calendar encoded at QP = 5 and transcoded at QP = 11 [8].

Based on the comparison of the various transcoding techniques, it can be inferred that from the point of view of low complexity and fast execution, pixel domain transcoding with motion vector reuse offers the best performance in terms of PSNR, execution time and resistance to drift. It may be mentioned that Lefol, Bull and Canagarajah [14] have evaluated the performance of transcoding algorithm for H.264. They have concluded that the fast pixel domain transcoder (FPDT) developed for MPEG-2 cannot be used for H.264 transcoding as it introduces unacceptable level of drift. The main contributions of this paper are 1) a standard deviation based decision mode for I frames, 2) the refinement and re-use of motion-vector information for P-frames, 3) the refinement and re-use of motion vector information for B-frames. These are described in detail in the next section.

## 4. PROPOSED TRANSCODING ARCHITECTURE AND TEST RESULTS

The proposed transcoding from MPEG-2 main profile to H.264 main profile is categorized into I, P, and B frames.

### 4.1. TRANSCODING AN I FRAME

H.264 performs adaptive spatial prediction to exploit the spatial redundancy in the I frame. It supports nine prediction modes for a 4×4 luma (luminance: black and white) sub-block and four prediction modes for a 16×16 luma macroblock. For chroma (color components) 4 different modes are defined, which are similar to the 4 modes for 16×16 intra luma prediction. Both the chroma blocks Cb and Cr use the same prediction mode. If a sub-block or a macroblock is to be coded in the intra mode, a prediction block is formed based on the neighboring samples of the previously coded blocks that are to the left and/or immediately above the block to be coded. The prediction block is then subtracted from the original macroblock to obtain the error residual.

The decision between the two modes (4×4 sub-block or 16×16 MB (macroblock)) (Fig. 9) relates to a tradeoff between compression and coding efficiency. When the MB contains intricate detailed information, it will be necessary to encode with maximum coding efficiency, otherwise it is necessary to exploit spatial redundancy and achieve maximum compression. A good measure of the information content of the macroblock is its information content.

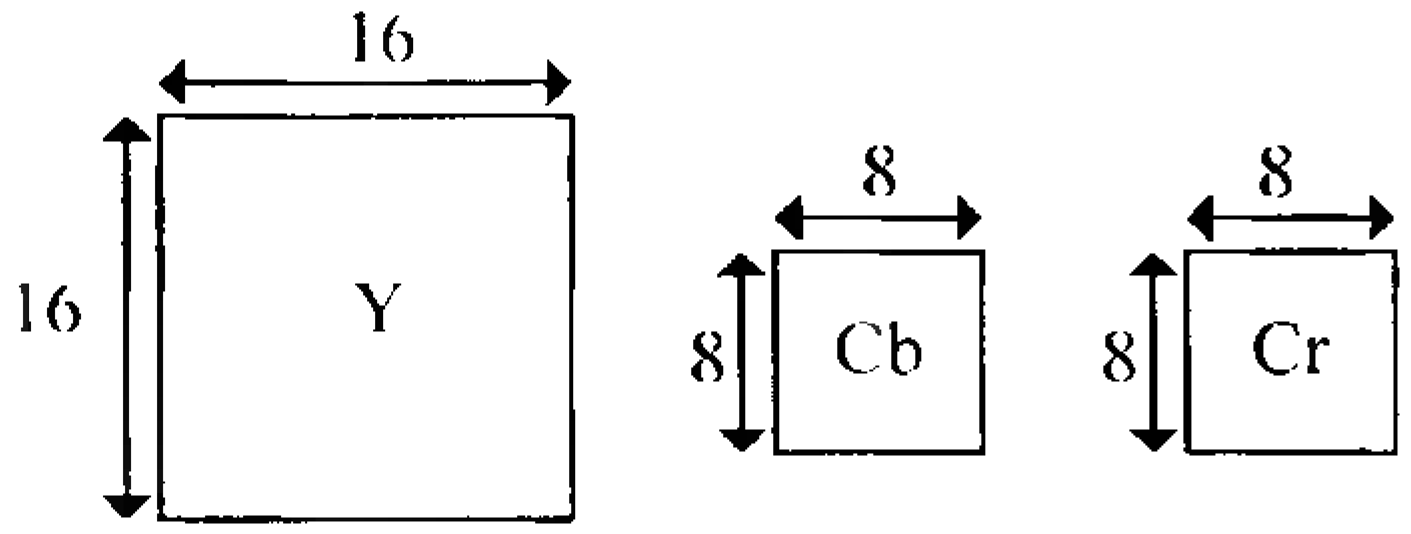


Fig. 9. (16×16) Macroblock for a 4:2:0 color format.

### *4.1.1. Experiments*

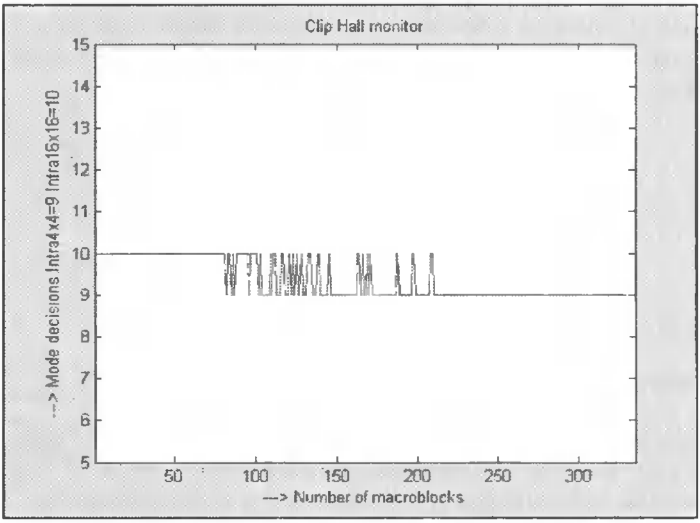
To compare the way the encoder chooses between these two intra modes and whether it has any relation with the statistical properties of the DCT coefficients like standard deviation, one must run several tests and plot the modes chosen versus the standard deviation of each macroblock in the I frame. The results obtained are shown in Fig. 10 for I frame of the test clip called "Hall Monitor". Similar characteristics hold for "Football" and "Akiyo" I frames. These are 352×288 progressive sequences (non-interlaced) which were encoded at a constant bit rate of 1 Mbps without rate-distortion optimization. As it can be seen, the macroblocks with a lower standard deviation tend to choose the 16×16 mode. After the first threshold, they tend to switch between the 16×16 mode and the 4×4 mode, and after the next higher threshold, they very distinctly choose the 4×4 mode. There are aberrations to this relationship, as can be observed. However, they are negligibly small. If this relation can be exploited during transcoding, since the DCT coefficients of the error residual are already available, the complexity of the intra-predictor block and mode decision can be completely avoided.

### *4.1.2. Proposed transcoder for I frames*

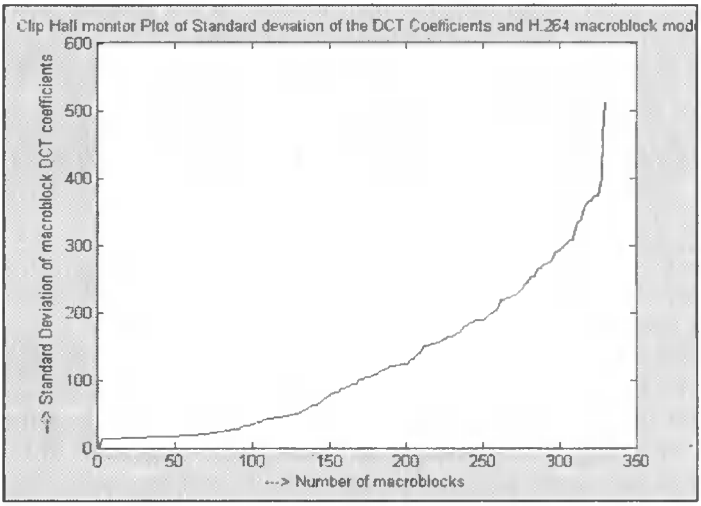
The general block diagram of the transcoder, which would implement standard deviation based I frame mode decisions is shown in Fig. 11.

The detailed flowchart of how a simple intra macroblock will be handled in software is shown in Fig. 12. The output from the decoder is the prediction residual obtained after the inverse DCT. The transform coefficients are used to make intra mode decisions. After adding up the residual, to obtain the reconstructed intra frame, it is directly fed into the encoder side, along with the mode decisions indicating whether the 16×16 or the 4×4 mode is to be used. Each macroblock is then predicted using these mode decisions. Sub-macroblock directional modes are chosen, based on macroblock availability and the sum of absolute difference value. For instance, suppose the 16×16 mode was selected. Then the availability of the surrounding macroblocks is determined. If only the macroblock to the left of the current macroblock is available then the 16×16-horizontal mode is selected as a valid mode for prediction. The sum of absolute difference (cost) for that mode is computed. The predicted macroblock is subtracted from the input macroblock and then the difference is transformed, quantized and entropy coded.

If there is more than one macroblock available and there is more than one valid mode, then the best mode selected is the one with the minimum cost.



(a) The mode decisions (intra 4×4 or 16×16) computed for an I frame in clip Hall Monitor vs. the number of macroblocks



(b) The corresponding values of the standard deviation of the macroblocks versus the number of macroblocks

Fig. 10. Results for I frame test clip Hall Monitor.

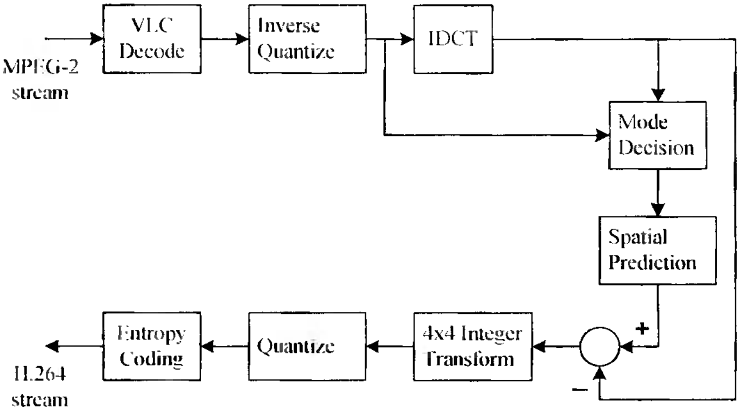


Fig. 11. Block diagram of the transcoder processing for an I frame.

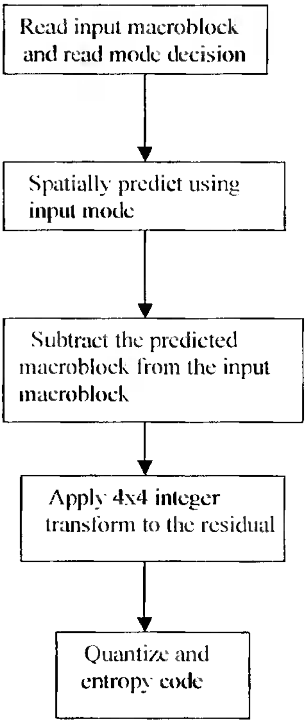

Read input macroblock and read mode decision
Spatially predict using input mode
Subtract the predicted macroblock from the input macroblock
Apply 4x4 integer transform to the residual
Quantize and entropy code

Fig. 12. Intra coding using previously computed variance based mode decisions.

### *4.1.3. Test results*

As shown in Fig. 13, the mode decision algorithm essentially captures the need for choosing coding efficiency over reduction in bit rate. It uses the standard deviation as an indication of the amount of information contained in the macroblock. It studies the pattern of the variations in the standard deviation within the blocks of a macroblock and determines the need to use either the intra-16×16-mode or intra-4×4-mode.

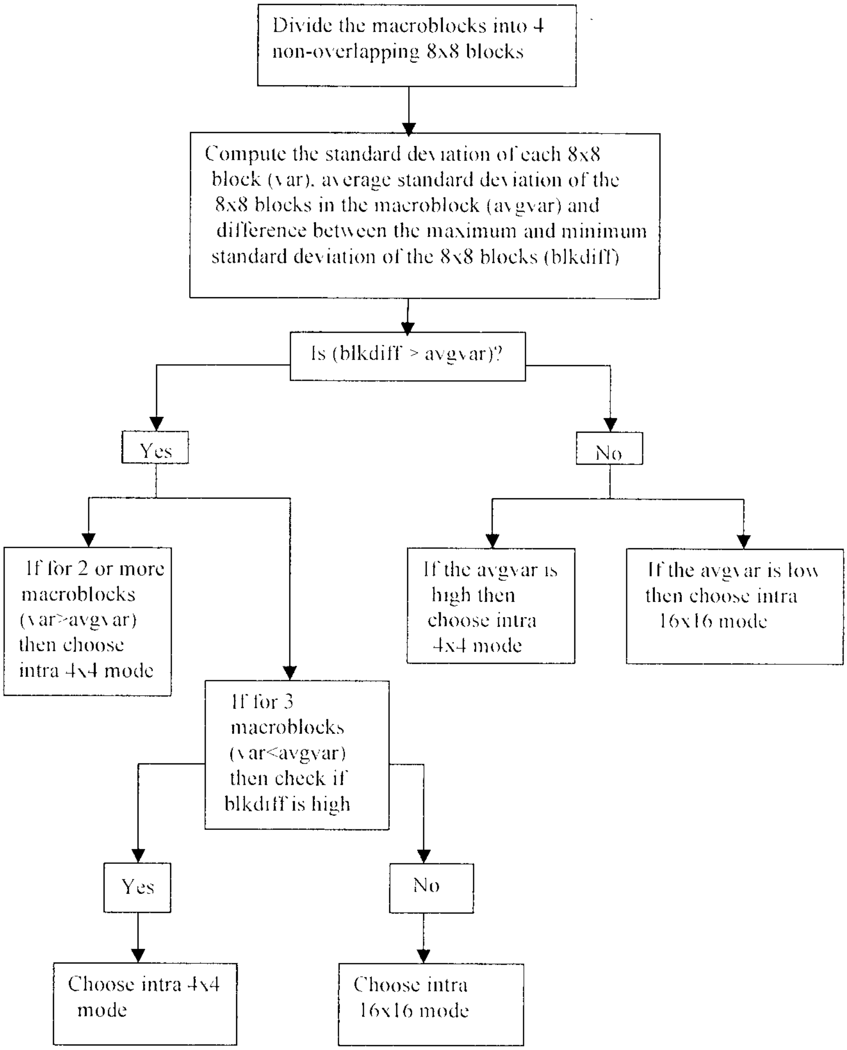

Divide the macroblocks into 4 non-overlapping 8x8 blocks
Compute the standard deviation of each 8x8 block (var), average standard deviation of the 8x8 blocks in the macroblock (avgvar) and difference between the maximum and minimum standard deviation of the 8x8 blocks (blkdiff)
Is (blkdiff > avgvar)?
Yes
No
If for 2 or more macroblocks (var>avgvar) then choose intra 4x4 mode
If for 3 macroblocks (var<avgvar) then check if blkdiff is high
If the avgvar is high then choose intra 4x4 mode
If the avgvar is low then choose intra 16x16 mode
Yes
No
Choose intra 4x4 mode
Choose intra 16x16 mode

Fig. 13. Mode decision algorithm for I frame transcoding.

Table 1 lists the results obtained for an I frame when the mentioned test clips were transcoded at 1 Mbps from an MPEG-2 bit stream to an H.264 bit stream. The mode decisions for the H.264 stream are chosen based on the algorithm described in Fig. 13. The PSNR (Fig. 14) obtained by the proposed new low complexity transcoding scheme (Table 1 and Fig. 15) is comparable to that obtained by complete decoding and re-encoding of the bit stream. The bits used, however, tend to be slightly higher than those in the case of full decoding and re-encoding. The full decoding and re-encoding process, performs an exhaustive search to find the best mode with the minimum cost. Hence the prediction residual is also very small and it requires less number of bits to be transmitted. However, in the proposed scheme, achieving low complexity (Table 1 and Fig. 15) is also of significant importance and hence a small increase in the number of bits used is acceptable with negligible loss in PSNR (Fig. 14). Figure 16 shows an I frame in the MPEG-2 stream, after transcoding and with complete decoding and re-encoding. As can be observed, the subjective quality of this frame based on the proposed transcoding (Fig. 16(b)) is similar to the later (Fig. 16(c)). Similar variance based mode decision (4×4 or 16×16) using QP as threshold has been explored in [15] for intra MB mode decision for MPEG-2 to H.264 transcoding. (QP is the quantization parameter used in quantizing the transform coefficients.) The proposed method, however, is independent of QP, and includes directional modes for both (4×4) and (16×16) blocks.

Table 1. Results obtained for the first I frame when the test clips were transcoded with variance based method decision algorithm.

| | PSNR (dB) (see Fig. 14) | | Bits used (bits/picture) | | Execution time (ms) (see Fig. 15) | |
|---|---|---|---|---|---|---|
| Test clip | Exhaustive SAD based mode selection | Standard deviation based mode selection | Exhaustive SAD based mode selection | Standard deviation based mode selection | Exhaustive SAD based mode selection | Standard deviation based mode selection |
| Akiyo | 43.221 | 43.086 | 56.480 | 67.720 | 300 | 240 |
| Coast guard | 39.798 | 39.81 | 119.880 | 126.024 | 420 | 331 |
| Football | 39.316 | 39.368 | 139.576 | 142.520 | 420 | 391 |
| Foreman | 41.448 | 41.4 | 75.008 | 92.520 | 301 | 241 |
| Crawfish | 42.2 | 42.153 | 68.864 | 77.592 | 301 | 240 |
| Flower garden | 39.232 | 39.303 | 215.200 | 217.376 | 381 | 311 |
| Hall monitor | 42.131 | 42.083 | 76.712 | 91.224 | 310 | 240 |

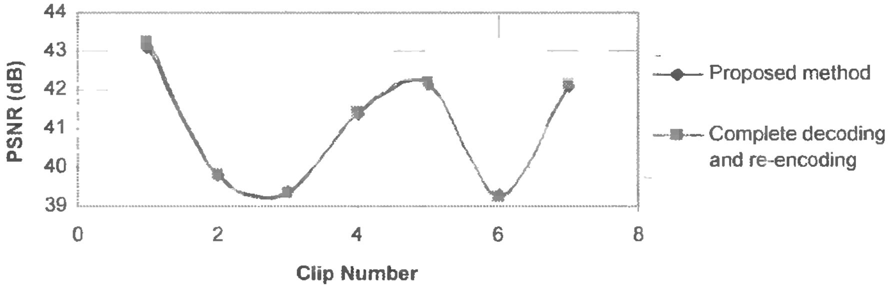


Fig. 14. Comparison of the PSNR of the proposed method and complete decoding and re-encoding method.

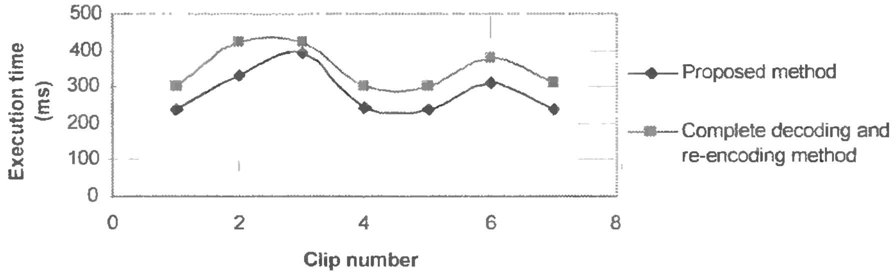


Fig. 15. Comparison of the execution time of the proposed method and complete decoding and re-encoding method.

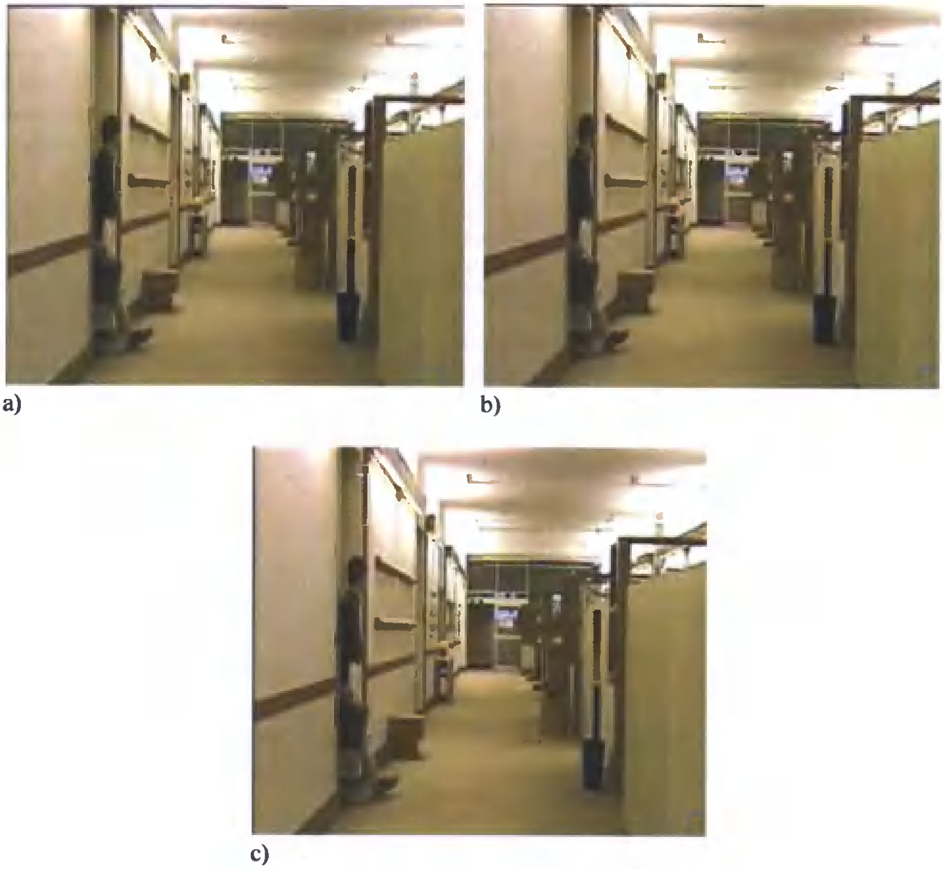


Fig. 16. Results of transcoded frames. Subjective quality of an I frame of the clip Hall Monitor in: a) an MPEG-2 compressed stream, b) the H.264 proposed transcoded compressed stream, c) the H.264 compressed stream.

## 4.2. P FRAME TRANSCODING

In the MPEG-2 standard [1], macroblocks in P frames can be coded alternatively using inter modes or the intra modes. Macroblocks that are inter-coded seek to exploit temporal correlation among frames and thus achieve compression. In P pictures, some modes available are MC/no-MC, coded/not-coded, intra/inter, and "quantizer modification" or not. The standard itself does not specify how to make these decisions.

The "MC/no-MC" decision is performed by the encoder, whether to transmit motion vectors or not. If the motion vector is zero due to MC decision then some bits can be saved by not transmitting it. The "intra/non-intra coding" decision is made based on variance. The "coded/not-coded" decision is a result of quantization. If all the quantized DCT coefficients are zero then the block need not be coded. The "quant/no-quant" decision is made as to whether the quantizer scale is to be changed or not. This is usually based on the frame content and on the decoder buffer status. The decision process for each macroblock can be described as shown in Fig. 17.

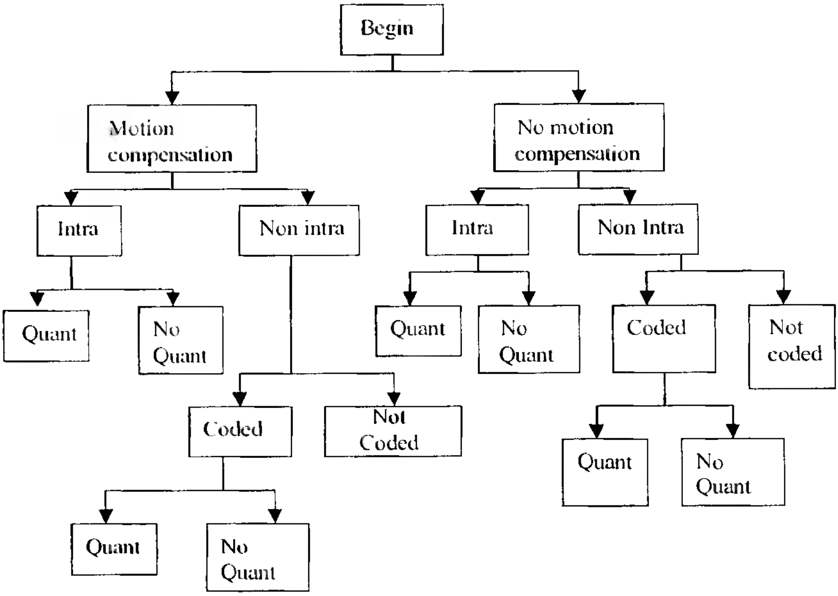


Fig. 17. Macroblock mode selection process for P frames in MPEG-2.

### *4.2.1. Coding mode decision in P frame*

Coding mode decisions are made by comparing sum of absolute difference value, i.e., motion cost for the refined motion vectors and the motion vectors predicted from the surrounding macroblocks in the same frame.

The motion cost for each pixel in a one pixel window (Fig. 18) is determined and the best one is selected based on the lowest cost as the best motion vector. To decide the best block size mode, top down splitting procedure is used as shown in Fig. 19. Using this top-down block splitting approach, initially the motion costs for the 16×16 block, the 8×16 blocks and the 16×8 blocks are determined. If the total motion cost for the 8×16 blocks or the 16×8 blocks exceeds the motion cost for the 16×16 block, then it implies that further partition may not significantly improve the performance. Hence the 8×8 and sub-8×8 block partitions are skipped from the motion estimation and mode decision process. Thus in general, if the costs for the next level of smaller sub blocks exceed that of the current level, further partitioning is stopped without checking the smaller blocks which come after the next level, to save computations. As shown in the Fig. 19, the same top-down approach is used for 8×8 blocks and sub-8×8 block partitions also. This helps to reduce the computational complexity of the coding mode decision (choice of block size for ME/MC).

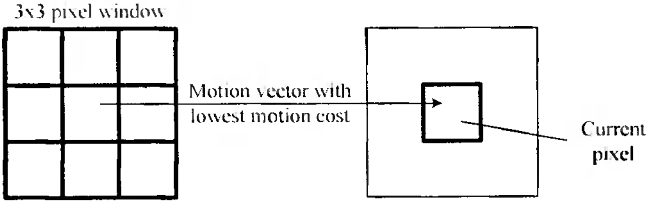


Fig. 18. Motion vector of a current pixel.

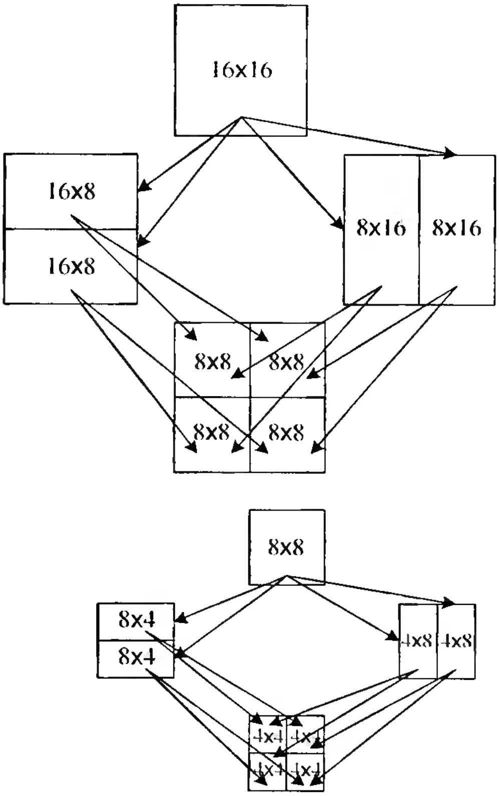


Fig. 19. Top down block splitting approach used to minimize the computational complexity of the encoding mode selection (choice of block size for ME/MC).

The overall scheme of P frame transcoding can be largely divided into two parts: extracting the frame data and motion vectors from the MPEG-2 bit stream, adjusting the motion vectors, refining them over a small window and reusing them in sub-pel motion estimation and mode decision.

### *4.2.2. Test results of P-frame transcoding*

The video clips used for this simulation are standard test clips [16], [17]. The objective of the proposed scheme for P frame transcoding is to reduce the complexity of the transcoding process (Table 2) without significantly changing the PSNR. It can be observed from Table 2, that the PSNR obtained by complete decoding and re-encoding of the MPEG-2 bit stream is comparable to that obtained by the proposed scheme. The PSNR also depends on other factors such as the motion search window size, bit rate, etc. For these experiments, the bit rate was maintained constant at 1 Mbps and the motion search window for the full re-encoding process was maintained at −1 pel to +1 pel and −24 pels to +24 pels. The savings in terms of execution time are also quite significant.

Note that the Figures 22(a)–23(c) indicate the motion vectors as marked and the mode decisions for the test clip "Akiyo" in all three scenarios: the original MPEG-2 bit stream, the transcoded H.264 bit stream and the H.264 bit stream obtained by complete decoding and re-encoding of the input. The mode decisions are indicated by the grid structure. In the case of MPEG-2 bit stream, since I frames support only 16×16 block motion compensated transform coding, the grid indicates the 16×16 block squares. In the case of the H.264 bit stream, the cells of the grid are sub-divided to indicate the kind of sub-macroblock partition, if any.

Table 2. The results obtained by transcoding a P frame from MPEG-2 to H.264 at 1 Mbps, compared with the complete MPEG-2 decoding and H.264 re-encoding of the MPEG-2 bit stream (see Figs. 20 and 21).

| | P frame with motion vector reuse and hierarchical mode decisions | | | P frame with full motion search, complete decoding and re-encoding | | |
|---|---|---|---|---|---|---|
| Test Clip | PSNR (dB) | Number of Bits used | Motion estimation time (MET) (ms) | PSNR (dB) | Number of Bits used | Motion estimation time (MET) (ms) |
| Crawfish | 40.249 | 60.080 | 631 | 34.349 | 50.392 | 4.196 |
| Akiyo | 41.952 | 13.536 | 401 | 42.137 | 10.464 | 3.976 |
| Coast guard | 37.531 | 115.544 | 751 | 37.93 | 94.688 | 4.136 |
| Football | 37.888 | 83.440 | 691 | 37.89 | 90.112 | 4.407 |
| Foreman | 39.835 | 38.280 | 581 | 40.09 | 32.620 | 4.016 |
| Flower garden | 36.882 | 212.120 | 671 | 37.349 | 134.880 | 4.196 |
| Hall monitor | 40.46 | 32.264 | 350 | 40.517 | 29.936 | 3.916 |

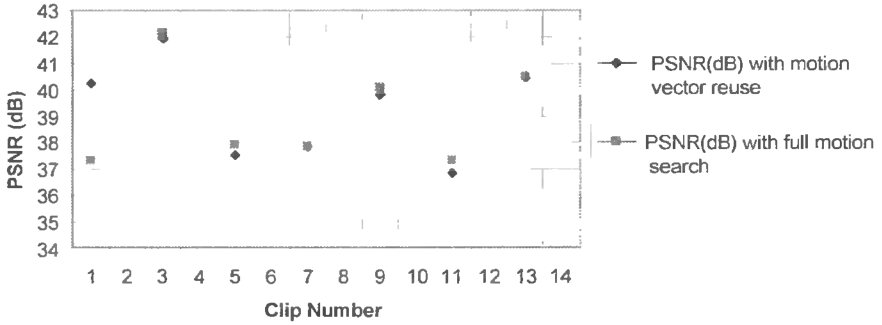


Fig. 20. Comparison of PSNR obtained by the proposed method of motion vector reuse vs. full motion search for P frames in different test clips.

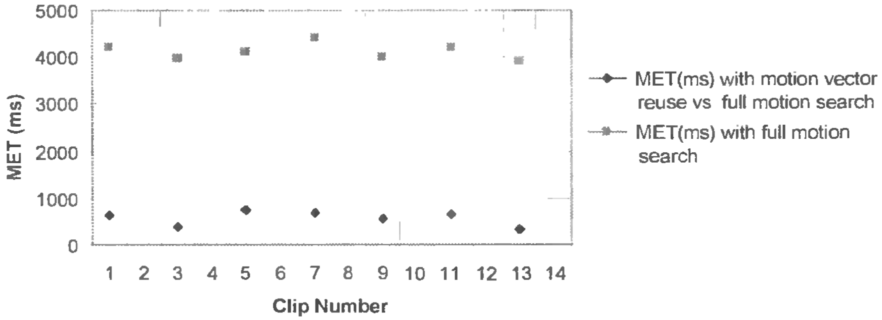


Fig. 21. Comparison of MET obtained by comparing the proposed method of motion vector reuse vs. full motion search for P frames in different test clips.

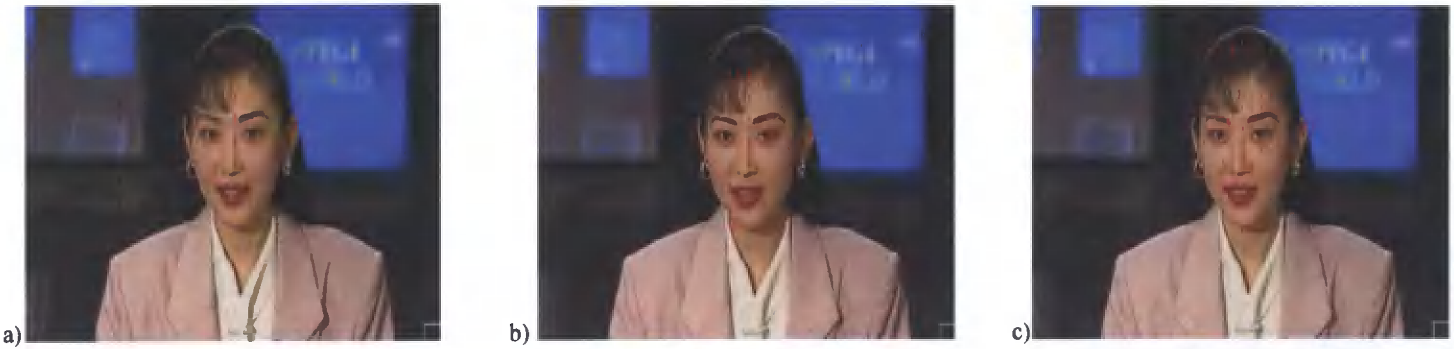
It can be observed that the results obtained by the proposed method are very close to that obtained by the full re-encoding process. For the test clip Akiyo, the motion vectors are shown in Fig. 22(a)–22(c) and the mode decisions can be observed in Fig. 23(a)–23(c). The mode decisions and the motion vectors computed are similar. The MPEG-2 motion vectors are coded for a 16×16 macroblock. Sub-macroblock partitions are not part of the MPEG-2 standard hence fewer motion vectors can be observed. The refinement scheme helps to take advantage of sub-macroblock modes and up to 50 percent more new motion vectors are determined. In this test clip, the motion is concentrated around the news reader's face, hence the detailed motion vectors and sub macroblock decisions can be observed around these areas. Similar observations can be made for the test clips listed in Table 2.

## 4.3. B FRAME TRANSCODING

The major difference between P frames and B frames is that there can be three types of prediction in B frames: forward, backward and interpolated prediction (Fig. 24). There are two estimated motion vectors (forward and backward) [2]. The motion vectors are set to zero at the start of each slice and at each intra macroblock.

### *4.3.1. B frame encoding in MPEG-2*

In the case of MPEG-2 [2], the encoder does not store B frames in the memory because they are not anchor frames. There are three ways of coding the motion vectors. If only a forward motion vector is present, then the motion compensated macroblock is predicted from the previous I or P picture. This is the same as the P frame where motion compensation is based on forward motion vectors only. If only a backward motion vector is present, the motion compensated macroblock is predicted from a future I or P picture. If both forward and backward motion vectors are present then the motion compensated macroblocks are constructed from both the previous I or P frame and the next I or P frame and the results of the two are averaged to form the interpolated motion compensated macroblock. As illustrated in Fig. 23, besides the intra mode and the predicted modes, the skip mode also exists for the B frames. However there exists one basic difference in the coding of skipped macroblock motion vectors for B frames. Skipped macroblocks in P frames have motion vectors equal to zero; however, skipped macroblocks in B frames have the motion vector differential equal to zero, which implies that the motion vectors of the current macroblock are the same as that of the previous macroblock. The skipped macroblock in a B frame has no VLC (variable length coding). The sequential mode decision procedures (Fig. 23) lead to macroblock type selection. The encoder first determines the motion compensation mode, i.e., forward, backward or interpolative. The "MC/no-MC" decision is not necessary for a B frame. The macroblocks are either regarded as motion compensated or as intra macroblock. The second decision is "intra or non-intra" coding. The intra macroblock is coded similarly as in the I frame. For the non-intra case, the prediction error is checked to see if it is large enough to be coded using the DCT. The last step is to decide whether the quantizer scale is satisfactory for the quantization of the DCT coefficients.

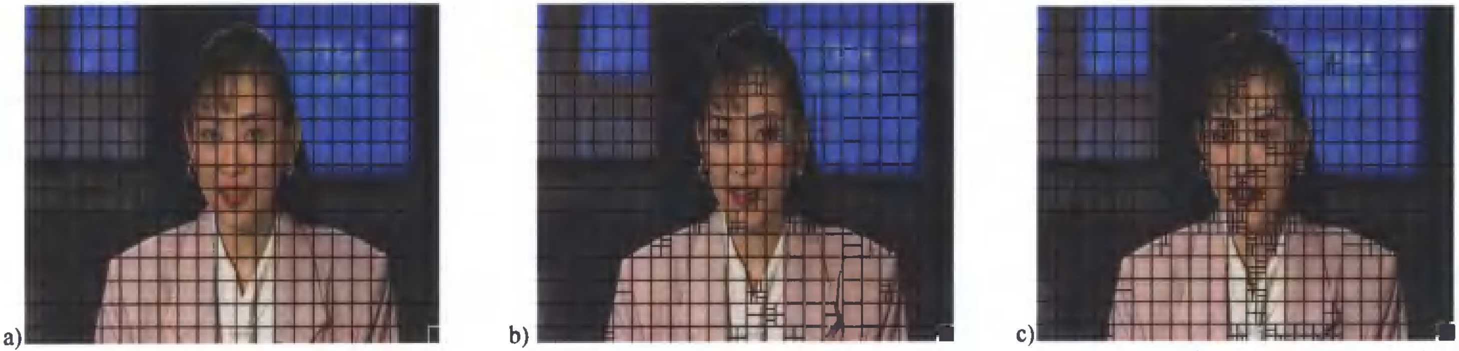


a) Shows a P frame of the 1 Mbps compressed MPEG-2 bit stream Akiyo with the motion vectors marked

b) Shows a P frame of the 1 Mbps compressed transcoded (proposed) H.264 bit stream Akiyo with the motion vectors marked

c) Shows a P frame of the 1 Mbps compressed H.264 bit stream Akiyo obtained by complete decoding and re-encoding with the motion vectors marked

Fig. 22. Results of transcoded P frames with motion vectors.



a) Shows a P frame of the 1 Mbps compressed MPEG-2 bit stream Akiyo with the mode decisions marked

b) Shows a P frame of the 1 Mbps compressed transcoded (proposed) H.264 bit stream Akiyo with the mode decisions marked

c) Shows a P frame of the 1 Mbps compressed H.264 bit stream Akiyo obtained by complete decoding and re-encoding with the mode decisions marked

Fig. 23. Results of transcoded P frames with motion vectors.

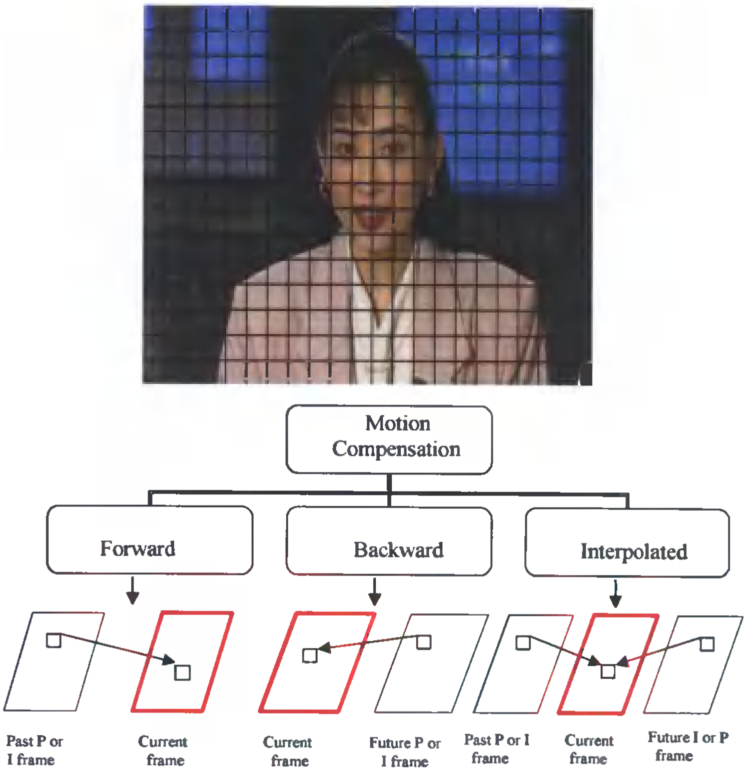


Fig. 24. Different coding modes for each macroblock in a B frame.

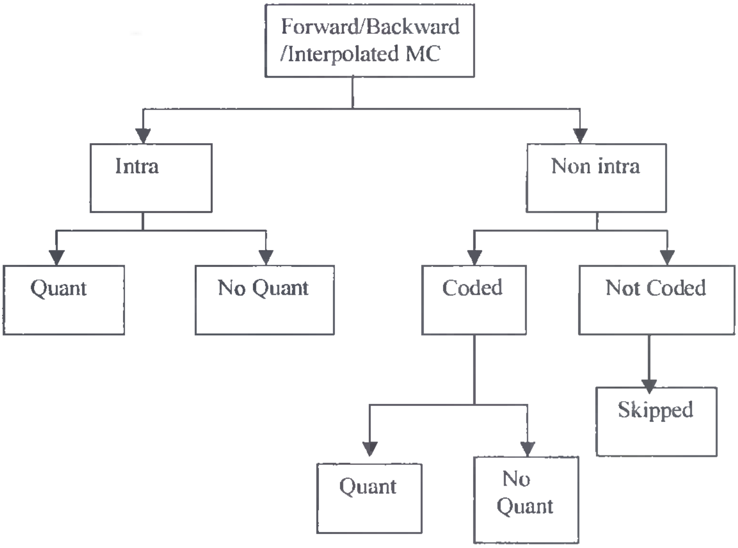


Fig. 25. Mode decision tree structure for macroblocks in B frame.

Blocks of 8×8 pixels are transformed into an array of 8×8 transform coefficients using the 2-D DCT that is the same as that for I and P pictures. Quantization and coding of the DCT coefficients in B pictures is the same as in P pictures.

### *4.3.2. B frame encoding in H.264*

B frame encoding in H.264/AVC is similar to that of MPEG-2; however there are certain additional capabilities in it that surpass MPEG-2. H.264 supports motion estimation and compensation for sub-macroblock partitions. For B frame macroblocks, each sub-partition can have up to two motion vectors allowed for temporal prediction [3]. They can be from any picture in the future or the past in display order. Hence H.264 supports multi-frame referencing which MPEG-2 does not support. There is a constraint on the maximum number of reference frames that can be used for prediction based on the profile and level being used [3]. Also H.264 allows the use of B frames as reference frames for temporal prediction.

As compared to MPEG-2, B frames also have a special mode in H.264 called the direct mode (Fig. 26). In this mode the motion vectors are not explicitly derived. The receiver obtains the motion vectors by scaling the motion vectors of a collocated macroblock in another reference picture. In this case, the reference picture for the current macroblock is the same as that of the collocated macroblock.

The weighted prediction concept [18] is further extended in the case of B frames. Weighted prediction can be used to enable the encoder adjustment of the weighting used in the weighted average between the two predictions that apply to bi-prediction. This can be especially effective for implementing "cross-fades" between two different video scenes, as bi-prediction allows flexible blending of content from these two scenes. The rest of the processing of B frames in H.264 remains the same as in the case of P frames.

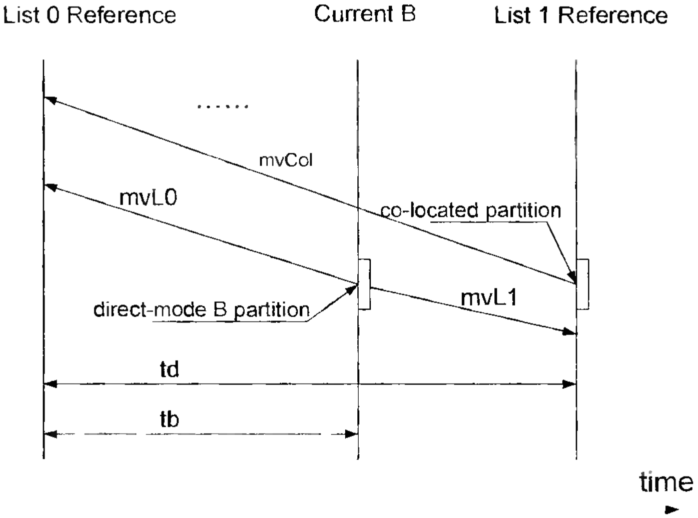


Fig. 26. Computation of motion vectors from the collocated macroblock for the direct node in B frames [3].

### *4.3.3. Transcoding of B frames*

Transcoding of B frames also requires motion vectors' reuse and refinement [7], [12], [19], as in the case of P frames. The proposed method refines them over a −1to +1 pixel window (Fig. 18) around the current pixel pointed to by the motion vectors. This process is repeated for the forward motion vector, as well as, for the backward motion vector. The search window size around the current pixel defines the accuracy of the refinement.

### *4.3.4. Test results*

Several tests were executed using different search window sizes and the results are as shown in Fig. 27. It can be observed that as the search window size is increased, initially the PSNR increases, however the increase tends to saturate to a steady state value above a certain search window size. It can also be noted that the one-pixel-window search provides a value close to the steady state PSNR value. Also, it involves the use of nine search centers which gives a fairly good tradeoff between computational complexity and PSNR. Similar results were obtained, when other test sequences were also tested. Hence a −1to +1 pixel window was selected as a reasonable tradeoff.

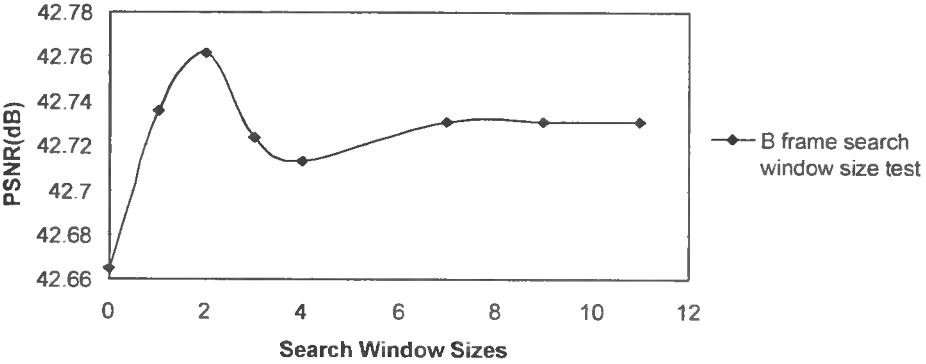


Fig. 27. Effect of the choice of search window sizes on PSNR, when tested on the clip Akiyo transcoded from a 1 Mbps MPEG-2 stream to a 699 kbps H.264 stream.

Further, as in the case of P frames the use of hierarchical mode decisions was also tested for different test sequences. The results for test sequence "Akiyo" are as tabulated in Tables 3 and 4, Figures 28 and 29. It can be observed that the PSNR in the two cases is almost the same, however; the execution time for the proposed method reduces by approximately 50% with the use of hierarchical mode decisions (see Fig. 29).

Table 3. Comparison of the PSNR, bit rate and execution time for the test sequence Akiyo when transcoded from a 1 Mbps MPEG-2 stream to an H.264 stream using the proposed method.

| Akiyo | | |
|---|---|---|
| PSNR (dB) | Bit rate (kbps) | Execution time (ms) |
| 32 | 137.46 | 2422 |
| 35.34 | 224.23 | 2815 |
| 38.58 | 379.86 | 2752 |
| 42.58 | 695.34 | 1070 |
| 45.31 | 1044.62 | 1071 |
| 49.11 | 1685.9 | 1062 |
| 52.93 | 2534.16 | 942 |

Table 4. Comparison of the PSNR, bit rate and execution time for the test sequence Akiyo when transcoded from a 1 Mbps MPEG-2 stream to an H.264 stream using the proposed method without hierarchical mode decision.

| Akiyo w/o hierarchical mode decision | | |
|---|---|---|
| PSNR (dB) | Bit rate (kbps) | Execution time (ms) |
| 32 | 137.46 | 4674 |
| 35.36 | 224.94 | 5426 |
| 38.59 | 380.37 | 5750 |
| 42.6 | 699.18 | 2434 |
| 45.33 | 1050 | 2463 |
| 49.14 | 1705.81 | 2353 |
| 52.95 | 2538.77 | 2352 |

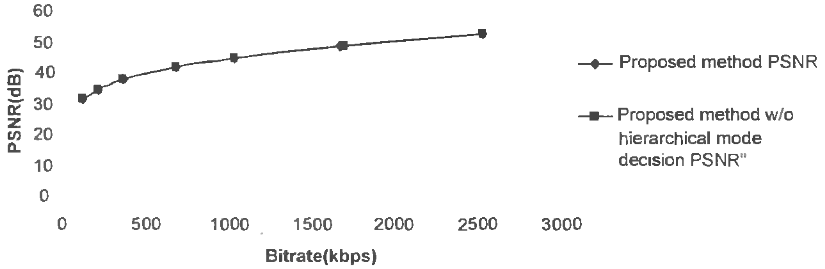


Fig. 28. Comparison of the PSNR in dB for the proposed reuse method with and without hierarchical mode decision.

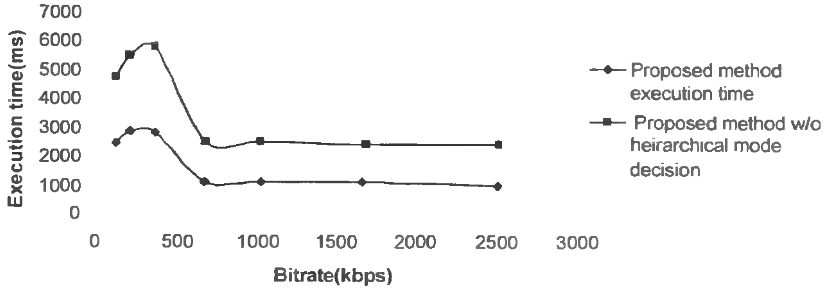
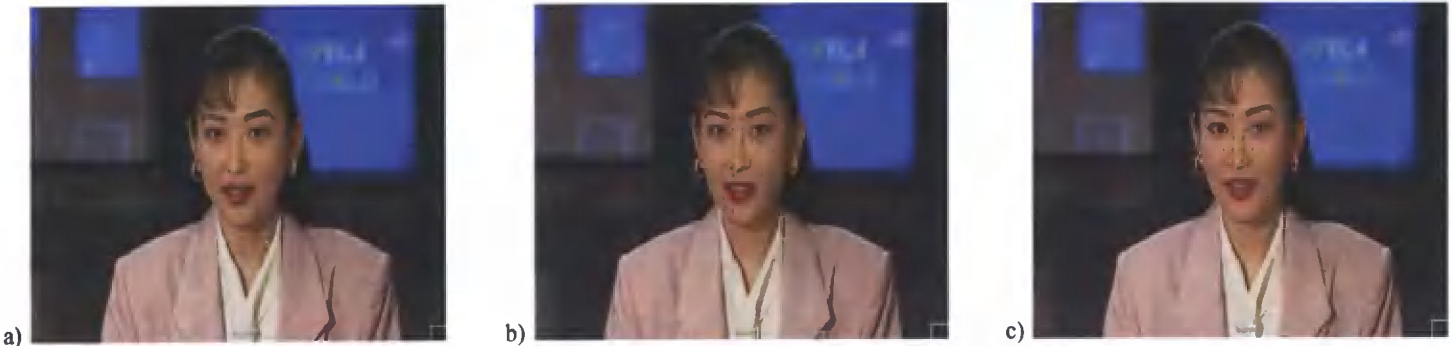


Fig. 29. Comparison of the execution time in ms for the methods as in Fig. 28.

Table 5. Comparison of the results obtained by transcoding different test sequences from a 1 Mbps MPEG-2 stream to an H.264 stream at a lower bit rate with those obtained by complete decoding and re-encoding of the same MPEG-2 stream.

| | B frame with motion vector reuse and hierarchical mode decisions | | | B frame with full motion search, complete decoding and re-encoding | | |
|---|---|---|---|---|---|---|
| Test Clip | PSNR (dB) | Number of Bits used | Motion Estimation Time (MET) (ms) | PSNR (dB) | Number of Bits used | Motion Estimation Time (MET) (ms) |
| Crawfish | 40.466 | 55.032 | 881 | 40.830 | 43.248 | 8.112 |
| Akiyo | 42.712 | 5.416 | 660 | 42.769 | 4.456 | 7.961 |
| Coast guard | 37.534 | 95.432 | 872 | 38.035 | 66.912 | 7.801 |
| Football | 37.922 | 74.696 | 801 | 37.905 | 71.776 | 7.993 |
| Foreman | 40.369 | 27.768 | 881 | 40.545 | 21.752 | 7.800 |
| Flower garden | 36.854 | 179.328 | 771 | 37.202 | 105.664 | 8.013 |
| Hall monitor | 41.175 | 15.528 | 660 | 41.210 | 13.368 | 7.482 |

The overall results while transcoding a B frame for different test clips are tabulated in Table 5. It can be observed that the proposed method has achieved 85–90% reduction in motion estimation complexity (compared to full motion search, complete decoding and re-encoding) with comparable PSNR values. The motion vectors and mode decisions for each of these test sequences can be observed in Figs. 30(a)–31(c).

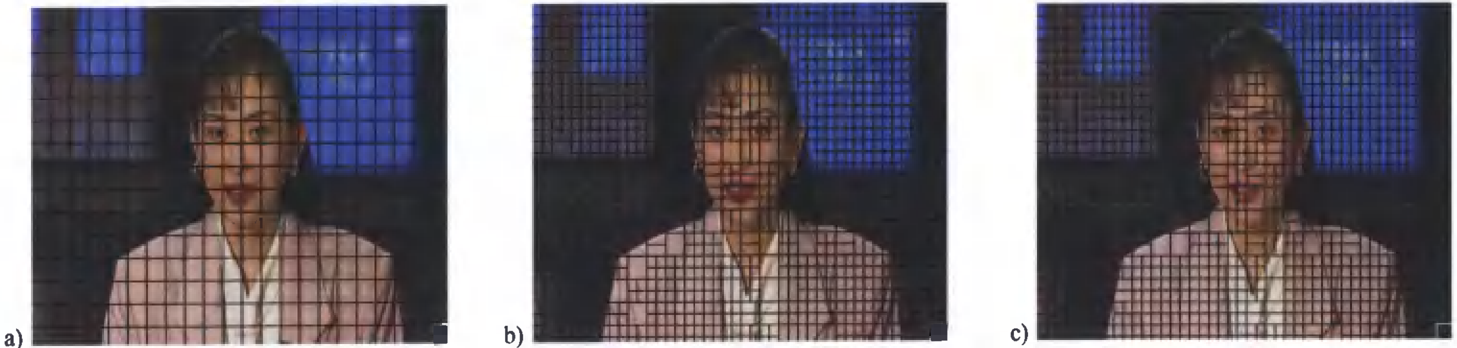


a) Shows a B frame of the 1 Mbps compressed MPEG-2 bit stream Akiyo with the forward motion vectors and backward motion vectors marked

b) Shows a B frame of the 1 Mbps compressed transcoded (proposed) H.264 bit stream Akiyo with the forward motion vectors and the backward motion vectors marked

c) Shows a B frame of the 1 Mbps compressed H.264 bit stream Akiyo obtained by complete decoding and re-encoding with forward motion vectors and backward motion vectors marked

Fig. 30. Results of transcoded B frames with forward and backward motion vectors.



a) Shows a B frame of the 1 Mbps compressed MPEG-2 bit stream Akiyo with the mode decisions marked

b) Shows a B frame of the 1 Mbps compressed transcoded (proposed) H.264 bit stream Akiyo with the mode decisions marked

c) Shows a B frame of the 1 Mbps compressed H.264 bit stream Akiyo obtained by complete decoding and re-encoding with the mode decisions marked

Fig. 31. Results of transcoded B frames with mode decisions.

Table 6. Comparison of the average PSNR of the input MPEG-2 bit stream and the H.264 transcoded bit stream measured with reference to the original source test clip.

| Test Clip | PSNR of the MPEG-2 bit stream (dB) | PSNR of the H.264 transcoded bit stream (dB) |
|---|---|---|
| Flower | 30.15 | 27.11 |
| Crawfish | 36.95 | 33.10 |
| Football | 28.12 | 26.11 |
| Foreman | 37.46 | 34.13 |
| Akiyo | 43.68 | 41.04 |
| Coast | 32.20 | 28.63 |
| Hall monitor | 36.93 | 34.01 |

## 5. TEST OF THE COMPLETE TRANSCODER

The results presented so far compared the proposed transcoding method with complete decoding and re-encoding of the input MPEG-2 bit stream. The two methods are very close and comparable in terms of PSNR (peak signal-to-noise ratio). However, a significant reduction in the execution time is achieved by the proposed method.

### 5.1. COMPARISON WITH COMPLETE DECODING AND RE-ENCODING

The test clips used are standard (352×240) CIF (common intermediate format) resolution test clips (Fig. 32). They are encoded into MPEG-2 streams at a bit rate of 1 Mbps and with a GOP size of 12 and the IBBPBBP··· GOP structure (Fig. 7). These streams are transcoded to 1 Mbps H.264 streams with the same GOP structure and GOP size (Table 6). As can be observed in Table 6, transcoding results in a reduction of about 2 -6 dB in the PSNR of the input bit stream. However, perceptually this reduction is not significantly visible.

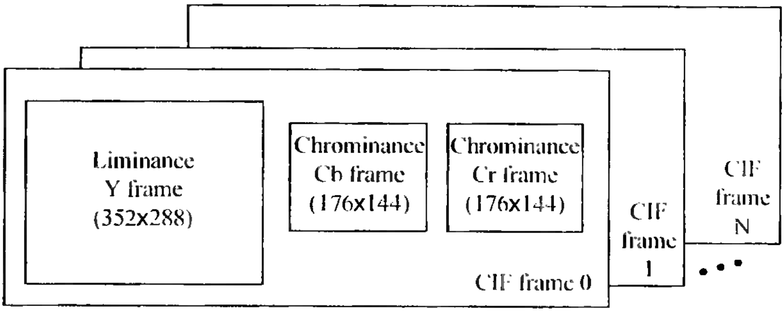


Fig. 32. Standard CIF (common intermediate format) frames in 4:2:0 sampling format of a test video clip.

Table 7. Comparison of the time (ms) to transcode a 1 Mbps input MPEG-2 bit stream Foreman to an H.264 bit stream at the same bit rate with the same GOP structure using PM, CDRE and DDT.

| | PM | DDT | CDRE |
|---|---|---|---|
| I frame | 681 | NA | 781 |
| P frame | 2.524 | 2.521 | 10.218 |
| B frame | 3.152 | 3.987 | 19.505 |

## 5.2. COMPARISON WITH OTHER STANDARDS

To derive conclusions about the performance of the proposed method it is very important to compare it with other proposed methods for transcoding. Figure 33 compares the proposed method (PM) with a DCT domain transcoder (DDT) [20] and with complete decoding and re-encoding (CDRE) of the MPEG-2 bit stream. The test is run on 30 frames of the test sequence Foreman at a bit rate of 1 Mbps and with a GOP structure of IBBPBBP···. The DCT domain transcoder [20] is shown in Fig. 34.

# 6. CONCLUSIONS

The proposed method clearly performs better than the complete decoding and re-encoding. Also, both the proposed method and CDRE perform better than transcoding in DCT domain. In terms of complexity, the average encoding time for each frame type is given in Table 7 for all the three methods.

The proposed method is very efficient in terms of encoding time. It is close and comparable to the DCT domain transcoder. However, CDRE is very computationally intensive as the method re-computes motion vectors, mode decisions, etc., without taking advantage of the data already present in the input MPEG-2 bit stream.

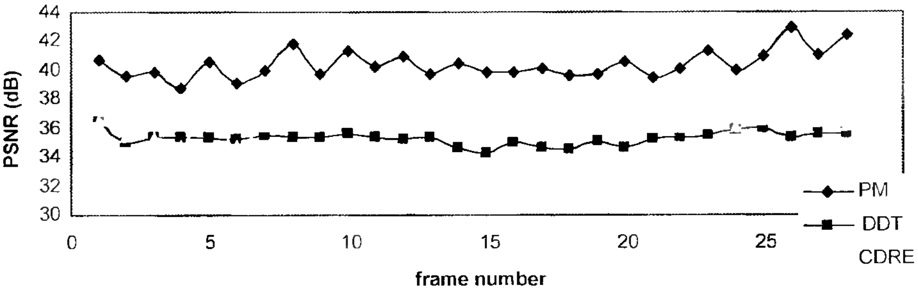


Fig. 33. Comparison of the proposed method (PM) with a DCT domain transcoder (DDT) and with complete decoding and re-encoding (CDRE) of the 1 Mbps MPEG-2 bit stream Foreman.

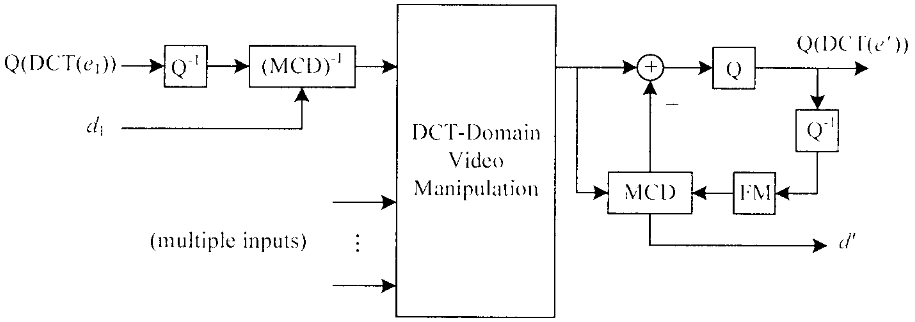


Fig. 34. DCT domain transcoder proposed by Chang and Messerschmitt [20]

## 7. FUTURE RESEARCH

The research presented in this paper is directed at low complexity, speed and comparable quality. As these targets have been achieved, the transcoder can be optimized for use in specific applications. For use in wireless environments, the major constraints would be adaptive network bandwidth usage, reduced spatial resolution and reduced frame rate. The proposed transcoder extracts the transform coefficients and auxiliary information and hence can be easily used to incorporate these requirements. Also, a strong error resilient rate control engine can be developed for the transcoder so that the bit rate of the transcoded stream can be varied as desired. Some scenarios require bit rate reduction transcoding techniques to accommodate changes in network bandwidth or variable bit rate transcoding for applications such as DVD recording.

## BIBLIOGRAPHY

[1] Information technology-generic coding of moving pictures and associated audio information: video, ITU-T and ISO/IEC JTC 1, ITU-T Rec. H.262 (2000E) and ISO/IEC 13818-2 (MPEG-2) Std.

[2] Rao K.R., Hwang J.J., 1996: Techniques and Standards for Image, Video and Audio Coding. Upper Saddle River, NJ: Prentice-Hall.

[3] Wiegand T., Sullivan G.J., Bjøntegaard G., Luthra A., 2003: Overview of the H.264/AVC video coding standard, IEEE Trans. Circuits Syst. Video Technol., vol. 13, 560-576.

[4] Sullivan G.J., Wiegand T., Luthra A., 2005: Draft of Version 4 of H.264 AVC, ITU-T Rec. H.264 and ISO/IEC 14496-10 (MPEG-4 part 10) Advanced Video Coding JVT Doc.

[5] Kwon S., Tamhankar A., Rao K.R., 2006: Overview of H.264/MPEG-4 part 10. J. Vis. Commun. Image R., vol. 17, pp. 186-216,.

[6] Sullivan G.J., 2005: The H.264/MPEG-4 AVC video coding standard and its deployment status, Proc. SPIE, Visual Commun. Image Process., vol. 5960. 709-719.

[7] Vetros A., Christopoulos C., Sun H., 2003: Video transcoding architectures and techniques: an overview, IEEE Signal Processing Mag., vol. 20. no. 2. 18-29.

[8] Xin J., Lin C.-W., Sun M.-T., 2005: Digital video transcoding, Proc. IEEE, vol. 93, 84-97.

[9] Sun H., Chiang T., Chen X., 2004: Digital Video Transcoding for Transmission and Storage. Boca Raton, FL: CRC Press.

[10] Kim J.-W. et al., 2006: Efficient video transcoding technique for QOS-based home gateway service, IEEE Trans. Consumer Electron., vol. 52, 129-137.

[11] McVeigh J. et al., 2000: A software based real-time MPEG-2 video encoder, IEEE Trans. Circuits Syst. Video Technol., vol. 10, pp. 1178–1184, Oct..

[12] Youn J., Sun M.-T., 1999: Motion vector refinement for high-performance transcoding," Proc. Int. Sym. Intelligent Multimedia, Video & Speech Processing, vol. 1, no. 1, 30-40.

[13] Wang J. et al., 2004: An AVS to MPEG-2 transcoding system, Proc. Int. Sym. Intelligent Multimedia, Video & Speech Processing, 302-305.

[14] Lefol D., Bull D., Canagarajah N., 2006: Performance evaluation of transcoding algorithms for H.264, IEEE Trans. Consumer Electron., vol. 52, 215-222.

[15] Petljanski B., Kalva H., 2006: DCT domain intra MB mode decision for MPEG-2 to H.264 transcoding, Proc. IEEE Int. conf. Consumer Electron., 419-420.

[16] Test bit streams and video clips. Tektronix Inc. [Online]. Available: ftp:// ftp.tek.com /tv/test/streams/Element/MPEG-Video/ 525/

[17] CIPR Sequences. Rensselaer Polytechnic Institute. Troy, NY. [Online]. Available: http:/ www.cipr.rpi.edu resource/ sequences/sif.html

[18] Puri A., Chen X., Luthra A., 2004: Video coding using the H.264/MPEG-4 AVC compression standard, Signal Processing: Image Communication, vol. 19, 787-937.

[19] Ghanbari M., 1999: Video Coding: an Introduction to Standard Codecs. London, UK: Institution of Electrical Engineers.

[20] Chang S.-F., Messerschmitt D.G., 1995: Manipulation and compositing of MC-DCT compressed video, IEEE J. Selected Areas Commun., vol. 13, 1-11.

## EFEKTYWNY SPOSÓB KODOWANIA STRUMIENIA BINARNEGO ZAPISANEGO W STANDARDZIE MPEG-2 DO STRUMIENIA BINARNEGO W STANDARDZIE H.264

### Streszczenie

Celem tej pracy jest przedstawienie opracowanej nowej metody transkodowania strumienia binarnego zapisanego w standardzie MPEG-2 na strumień binarny w standardzie H.264 oraz porównanie jej parametrów z innymi znanymi metodami. Opracowany nowy transkoder charakteryzuje się tym, że wykorzystuje wielokrotnie informację zawartą w strumieniu standardu MPEG-2 w zastosowanych w nim takich technikach jak: wieloblokowa estymacja ruchu, filtr deblokujący wbudowany w pętli, wewnętrzna predykcja kierunkowa, transformacja DCT oparta o liczby całkowite, adaptacyjne kontekstowe kodowanie o zmiennej długości, adaptacyjna predykcja z wagami oraz użycie wag opartych na ocenie ludzkiej wrażliwości wizualnej – zastosowanych w standardzie H.264. W zaproponowanej metodzie osiąga się niską złożoność oraz porównywalną jakość i zredukowaną przepływność w procesie transkodowania w porównaniu ze stosowanymi do tej pory technikami.

Słowa kluczowe: standardy MPEG-2 i H.264, transkodowanie wideo

UNIWERSYTET TECHNOLOGICZNO-PRZYRODNICZY
IM. JANA I JĘDRZEJA ŚNIADECKICH W BYDGOSZCZY
ZESZYTY NAUKOWE NR 250
TELEKOMUNIKACJA I ELEKTRONIKA 11 (2008), 33-52

# ONE-STEP 9-STAGE HERMITE-BIRKHOFF-TAYLOR ODE SOLVER OF ORDER 11

Vladan Bozic, Artur Przybylo, Truong Nguyen-Ba, Rémi Vaillancourt

University of Ottawa – Department of Mathematics and Statistics
585 King Edward Ave., Ottawa ON, K1N 6N6 Canada

*Summary*: A one-step 9-stage Hermite-Birkhoff-Taylor method of order 11, denoted by HBT(11)9, is constructed for solving nonstiff systems of first-order differential equations of the form $y' \doteq f(x,y)$, $y(x_0)=y_0$. The method uses $y'$ and the higher derivatives $y^{(2)}$ to $y^{(5)}$ as in Taylor methods and is combined with a 9-stage Runge-Kutta method. Forcing a Taylor expansion of the numerical solution to agree with an expansion of the true solution leads to Taylor- and Runge-Kutta-type order conditions which are reorganized into Vandermonde-type linear systems whose solutions are the coefficients of the method. The new method has larger scaled interval of absolute stability than Dormand-Prince DP(8,7)13M. The stepsize is controlled by means of $y^{(3)}$ and $y^{(5)}$. HBT(11)9 is superior to Dormand-Prince DP(8,7)13M and Taylor method of order 11 in solving several problems often used to test high-order ODE solvers on the basis of the number of steps, CPU time, and maximum global error. These numerical results show the benefits of adding high-order derivatives to Runge-Kutta methods.

Keywords: general linear method for non-stiff ODE's, Hermite-Birkhoff-Taylor method, maximum global error, number of function evaluations, CPU time.

## 1. INTRODUCTION

A Taylor method of order 5, denoted by T5, and a 9-stage Runge-Kutta method of order 7 are cast into a one-step 9-stage Hermite-Birkhoff-Taylor method of order 11, named HBT(11)9 because it uses Hermite-Birkhoff interpolation polynomials and the derivatives $y'$ and $y^{(2)}$ to $y^{(5)}$ for solving $y'=f(x,y)$ at step points. The link between the two types of methods is that values at off-step points are obtained by means of predictors which use values of derivatives of different orders at the current step point. By construction, HBT(11)9 uses lower order derivatives than the traditional Taylor method of order 11, denoted by T11 [11].

Taylor methods have been an excellent choice in astronomical calculations [3] and sensitivity analysis of ODEs/DAEs [2], and in solving general problems [5] and validating solutions of ODEs/DAEs by means of interval analysis [9], [12]. Deprit and Zahar [7] proved that recurrent power series in Taylor methods are very effective in achieving high accuracy, with less computing time and larger stepsize.

HBT(11)9 is designed for solving nonstiff systems of first-order initial value problems of the form

$$y' = f(x,y), \qquad y(x_0) = y_0, \qquad \text{where} \quad ' = \frac{d}{dx} \tag{1}$$

The high-order derivatives $y^{(2)}$ to $y^{(5)}$ can be obtained by differentiating $f(x,y(x))$ in the right-hand side of equation (1). But this approach is useful only in theoretical studies because of the computational complexity of high-order derivatives.

Following the pioneering work of Steffensen [20] and [18], another approach uses fast automatic differentiation (AD) techniques to compute sums, differences, products and powers of power series, to name but a few (see [3], [11], and references therein). Formulae for generating these high-order derivatives can be found in textbooks (see, for instance, [8]).

Forcing a Taylor expansion of the numerical solution to agree with an expansion of the true solution leads to a combination of Taylor- and Runge-Kutta-type order conditions which are reorganized into linear Vandermonde-type systems. The coefficients of this one-step method are the solutions of these systems and can be obtained by means, say, of Gaussian elimination. Moreover, with HBT(11)9, there are no rejected steps because the chosen stepsize yields the required precision level once the series is generated.

HBT(11)9 has larger scaled intervals of absolute stability than DP(8,7)13M. The C++ performances of HBT(11)9, DP(8,7)13M [17] and T11, were compared on several problems frequently used to test higher order ODE solvers. It is seen that, generally, HBT(11)9 requires fewer steps, uses less CPU time, and has higher accuracy than DP(8,7)13M and T11. Other HBT methods have been studied [13,14,15].

Section 2 introduces HBT(11)9. Order conditions are listed in Section 3. In Section 4, HBT(11)9 is represented in terms of Vandermonde-type systems. Section 5 considers the region of absolute stability of the method. Section 6 deals with the step control. In Section 7, two criteria are used to compare the performance of the methods considered in this paper. Appendix A lists the defining formulae of HBT(11)9 and Appendix B briefly describes the recurrent computation of higher-order derivatives.

## 2. ONE-STEP HBT(11)9

HBT(11)9 requires eight predictors, $P_2$, $P_3$, …, $P_9$, and an integration formula, IF, to perform the integration step from $x_n$ to $x_{n+1}$.

Hermite-Birkhoff polynomials of increasing degree are used as predictors $P_\ell$ to obtain $y_{n+c_\ell}$ to orders 5 for $\ell = 2$, order 6 for $\ell = 3$, and order 7 for $\ell = 4,\dots,9$, respectively,

$$y_{n+c_\ell} = y_n + h_{n+1}\sum_{j=1}^{\ell-1} a_{\ell j} f_{n+c_j} + \sum_{j=2}^{5} h_{n+1}^{j} \gamma_{\ell j} f_n^{(j-1)}, \qquad \ell = 2,3,\dots,9\,. \tag{2}$$

A Hermite-Birkhoff polynomial of degree 11 is used as integration formula IF to obtain $y_{n+1}$ to order 11,

$$y_{n+1} = y_n + h_{n+1}\sum_{j=1}^{9} b_j f_{n+c_j} + \sum_{j=2}^{5} h_{n+1}^{j}\gamma_{1j} f_n^{(j-1)} . \tag{3}$$

By notation, $y_{n+c_9}$ is different from $y_{n+1}$ when $c_9 = 1$. One sees that the derivatives $f_n^{(1)}$ to $f_n^{(4)}$ are computed only once per step at $x_n$. The recurrent computation of these high-order derivatives is briefly described in Appendix B.

The reader can see that the defining formulae of HBT(11)9 involve the Runge-Kutta parameters listed in the following Butcher tableau:

$$\begin{array}{c|ccccccccc}
c_1 & & & & & & & & & \\
c_2 & a_{21} & & & & & & & & \\
c_3 & a_{31} & a_{32} & & & & & & & \\
c_4 & a_{41} & a_{42} & a_{43} & & & & & & \\
c_5 & a_{51} & a_{52} & a_{53} & a_{54} & & & & & \\
c_6 & a_{61} & a_{62} & a_{63} & a_{64} & a_{65} & & & & \\
c_7 & a_{71} & a_{72} & a_{73} & a_{74} & a_{75} & a_{76} & & & \\
c_8 & a_{81} & a_{82} & a_{83} & a_{84} & a_{85} & a_{86} & a_{87} & & \\
c_9 & a_{91} & a_{92} & a_{93} & a_{94} & a_{95} & a_{96} & a_{97} & a_{98} & \\
\hline
 & b_1 & b_2 & b_3 & b_4 & b_5 & b_6 & b_7 & b_8 & b_9
\end{array}$$

and the Taylor expansion parameters $\gamma_{ij}$.

## 3. ORDER CONDITIONS FOR HBT(11)9

We impose the following simplifying assumptions on HBT(11)9, with $\gamma_{i1} = 0$ and $\gamma_{i,m} = 0$ for $m > 5$:

$$\sum_{i=j+1}^{9} b_i a_{ij} = b_j(1-c_j), \qquad j = 2,\ldots,8, \tag{4}$$

$$b_2 = b_3 = 0, \tag{5}$$

$$a_{i2} = 0, \qquad i = 4,\cdots,9, \tag{6}$$

$$\sum_{j=1}^{i-1} a_{ij}c_j^k + k!\gamma_{i,k+1} = \frac{1}{k+1}c_i^{k+1}, \qquad i = 2,3,\ldots,9, \quad k = 0,1,\ldots,4, \tag{7}$$

$$\sum_{j=1}^{i-1} a_{ij}c_j^5 + 5!\gamma_{i7} = \frac{1}{6}c_i^6, \qquad i = 3,\ldots,9, \tag{8}$$

$$\sum_{j=1}^{i-1} a_{ij}c_j^6 + 6!\gamma_{i8} = \frac{1}{7}c_i^7, \qquad i = 4,\ldots,9. \tag{9}$$

There remain seven sets of equations to be solved:

$$\sum_{i=1}^{9} b_i c_i^k + k!\gamma_{1,k+1} = \frac{1}{k+1}, \qquad k = 0,1,\ldots,10\,, \tag{10}$$

$$\begin{aligned}
&b_8(1-c_8)a_{87}c_7^5(c_7-c_4)(c_7-c_5)(c_7-c_6) = \\
&= 8!\left(\frac{1}{10!}-\frac{10}{11!}\right) - 7!\left(\frac{1}{9!}-\frac{9}{10!}\right)(c_4+c_5+c_6) + \\
&+6!\left(\frac{1}{8!}-\frac{8}{9!}\right)(c_4c_5+c_4c_6+c_5c_6) - 5!\left(\frac{1}{7!}-\frac{7}{8!}\right)c_4c_5c_6
\end{aligned} \tag{11}$$

$$\begin{aligned}
&b_7(1-c_7)(c_8-c_7)a_{76}c_6^5(c_6-c_4)(c_6-c_5) = \\
&= 7!\left[\frac{c_8}{9!}-(1+c_8)\frac{9}{10!}+9\frac{10}{11!}\right] - 6!\left[\frac{c_8}{8!}-(1+c_8)\frac{8}{9!}+8\frac{9}{10!}\right](c_4+c_5) + \\
&+5!\left[\frac{c_8}{7!}-(1+c_8)\frac{7}{8!}+7\frac{8}{9!}\right]c_4c_5
\end{aligned} \tag{12}$$

$$\begin{aligned}
&b_8(1-c_8)a_{87}c_7^5(c_7-c_4)(c_7-c_5) + b_8(1-c_8)a_{86}c_6^5(c_6-c_4)(c_6-c_5) + \\
&+b_7(1-c_7)a_{76}c_6^5(c_6-c_4)(c_6-c_5) = \\
&= \left(\frac{7!}{9!}-9\frac{7!}{10!}\right) - (c_4+c_5)\left(\frac{6!}{8!}-8\frac{6!}{9!}\right) + c_4c_5\left(\frac{5!}{7!}-7\frac{5!}{8!}\right)
\end{aligned} \tag{13}$$

$$\sum_{i=4}^{7} b_i(1-c_i)(c_8-c_i)a_{i3} = 0 \tag{14}$$

$$\sum_{i=4}^{8} b_i(1-c_i)a_{i3} = 0 \tag{15}$$

$$\sum_{i=5}^{8} b_i(1-c_i)\sum_{j=4}^{i-1} a_{ij}a_{j3} = 0 \tag{16}$$

The nine off-step points used in this paper are

$$\begin{aligned}
c_1 &= 0 \\
c_2 &= 0.27921731436446077438372923301803 \\
c_3 &= 0.32575353342520423678101743852104 \\
c_4 &= 0.38004578899607160957785367827455 \\
c_5 &= 0.62571583978357747657383924888563 \\
c_6 &= 0.68469945467088040000000000000000 \\
c_7 &= 0.79537958163786415255458450701553 \\
c_8 &= 0.92465627764050439818532822755515 \\
c_9 &= 1
\end{aligned} \tag{17}$$

The abscissae $c_5$ and $c_7$ are determined such that the Gauss-type integration formula with 5 multiple preassigned abscissae $c_1 = 0$ and three preassigned abscissae

$$c_4 = 0.38004578899607160957785367827455$$
$$c_8 = 0.92465627764050439818532822755515$$
$$c_9 = 1$$

is exact for polynomials of degree $\leq 11$ [6]. The values of $c_3$ and $c_2$ are $c_3 = (6/7)c_4$ and $c_2 = (6/7)c_3$.

We note that, in this paper, the coefficients of the method are given to 32 digits for increased accuracy at stringent tolerances and for use in extended precision computation.

To obtain the value $c_6$ of (17), firstly, we write the following reduced equation

$$\begin{aligned} &b_8(1-c_8)a_{87}a_{76}c_6^5(c_6-c_4)(c_6-c_5) = \\ &= 7!\left(\frac{1}{10!}-\frac{10}{11!}\right)-(c_4+c_5)6!\left(\frac{1}{9!}-\frac{9}{10!}\right)+c_4c_5 5!\left(\frac{1}{8!}-\frac{8}{9!}\right). \end{aligned} \quad (18)$$

Secondly, we write

$$\theta = c_7^5(c_7-c_4)(c_7-c_5)(c_7-c_6)b_7(1-c_7)(c_8-c_7),$$

so that the product of the left-hand sides of (11) and (12) is the product of $\theta$ with the left-hand side of (18). We therefore have

$$\begin{aligned} &\left[8!\left(\frac{1}{10!}-\frac{10}{11!}\right)-(c_4+c_5+c_6)7!\left(\frac{1}{9!}-\frac{9}{10!}\right)+\right. \\ &\left.+(c_4c_5+c_4c_6+c_5c_6)6!\left(\frac{1}{8!}-\frac{8}{9!}\right)-c_4c_5c_6 5!\left(\frac{1}{7!}-\frac{7}{8!}\right)\right]\times \\ &\times\left[7!\left(\frac{c_8}{9!}-(1+c_8)\frac{9}{10!}+9\frac{10}{11!}\right)-(c_4+c_5)6!\left(\frac{c_8}{8!}-(1+c_8)\frac{8}{9!}+8\frac{9}{10!}\right)+\right. \\ &\left.+c_4c_5 5!\left(\frac{c8}{7!}-(1+c_8)\frac{7}{8!}+7\frac{8}{9!}\right)\right]= \\ &=\left[7!\left(\frac{1}{10!}-\frac{10}{11!}\right)-(c_4+c_5)6!\left(\frac{1}{9!}-\frac{9}{10!}\right)+c_4c_5 5!\left(\frac{1}{8!}-\frac{8}{9!}\right)\right]\theta. \end{aligned} \quad (19)$$

Setting $c_i$ equal to the values of (17) for all $i$ except $i=6$, we can calculate $c_6$ such that (19) is satisfied.

## 4. VANDERMONDE-TYPE FORMULATION OF HBT(11)9

### 4.1. INTEGRATION FORMULA IF

The 11-vector of the reordered coefficients of IF in (3) is the solution of the Vandermonde-type system of order conditions:

$$\begin{bmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 \\ c_9 & c_8 & c_7 & c_6 & c_5 & c_4 & 0 & 1 & 0 & 0 & 0 \\ c_9^2/2! & c_8^2/2! & c_7^2/2! & c_6^2/2! & c_5^2/2! & c_4^2/2! & 0 & 0 & 1 & 0 & 0 \\ c_9^3/3! & c_8^3/3! & c_7^3/3! & c_6^3/3! & c_5^3/3! & c_4^3/3! & 0 & 0 & 0 & 1 & 0 \\ c_9^4/4! & c_8^4/4! & c_7^4/4! & c_6^4/4! & c_5^4/4! & c_4^4/4! & 0 & 0 & 0 & 0 & 1 \\ c_9^5/5! & c_8^5/5! & c_7^5/5! & c_6^5/5! & c_5^5/5! & c_4^5/5! & 0 & 0 & 0 & 0 & 0 \\ c_9^6/6! & c_8^6/6! & c_7^6/6! & c_6^6/6! & c_5^6/6! & c_4^6/6! & 0 & 0 & 0 & 0 & 0 \\ c_9^7/7! & c_8^7/7! & c_7^7/7! & c_6^7/7! & c_5^7/7! & c_4^7/7! & 0 & 0 & 0 & 0 & 0 \\ c_9^8/8! & c_8^8/8! & c_7^8/8! & c_6^8/8! & c_5^8/8! & c_4^8/8! & 0 & 0 & 0 & 0 & 0 \\ c_9^9/9! & c_8^9/9! & c_7^9/9! & c_6^9/9! & c_5^9/9! & c_4^9/9! & 0 & 0 & 0 & 0 & 0 \\ c_9^{10}/10! & c_8^{10}/10! & c_7^{10}/10! & c_6^{10}/10! & c_5^{10}/10! & c_4^{10}/10! & 0 & 0 & 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} b_9 \\ b_8 \\ b_7 \\ b_6 \\ b_5 \\ b_4 \\ b_1 \\ \gamma_{12} \\ \gamma_{13} \\ \gamma_{14} \\ \gamma_{15} \end{bmatrix} = \begin{bmatrix} 1 \\ 1/2! \\ 1/3! \\ 1/4! \\ 1/5! \\ 1/6! \\ 1/7! \\ 1/8! \\ 1/9! \\ 1/10! \\ 1/11! \end{bmatrix} \qquad (20)$$

With the choice of $c_i$, $i = 4,5,\ldots,9$, in (17), the leading error term of IF is of order 13:

$$\left[ b_9 \frac{c_9^{12}}{12!} + \cdots + b_5 \frac{c_5^{12}}{12!} + b_4 \frac{c_4^{12}}{12!} - \frac{1}{13!} \right] h_{n+1}^{13} y_n^{(13)}.$$

### 4.2. PREDICTOR $\mathrm{P}_2$

The $i$ th component, $u_2(i)$, of the 5-vector of reordered coefficients of predictor $\mathrm{P}_2$ in (2) with $\ell = 2$,

$$\mathbf{u}^2 = [a_{21}, \gamma_{22}, \gamma_{23}, \gamma_{24}, \gamma_{25}]^T,$$

satisfies the order condition

$$u_2(i) = \frac{c_2^i}{i!}, \qquad i = 1,2,\ldots,5. \qquad (21)$$

A truncated Taylor expansion of the right-hand side of (2) with $\ell = 2$ about $x_n$ gives

$$\sum_{j=0}^{12} S_2(j) h_{n+1}^j y_n^{(j)}$$

with coefficients

$$2S_2(j) = M^2(j,1:5)\mathbf{u}^2 = \frac{c_2^j}{j!}, \qquad j = 1,2,\ldots,5,$$

$$S_2(j) = 0, \qquad j = 6,7,\ldots,12.$$

We note that $\mathrm{P}_2$ is of order 5 since it satisfies the order conditions

$$S_2(j) = \frac{c_2^j}{j!}, \qquad j = 1,2,\ldots,5$$

and its leading error term is

$$\left[ S_2(6) - \frac{c_2^6}{6!} \right] h_{n+1}^6 y_n^{(6)}.$$

### 4.3. PREDICTOR P$_3$

The 6-vector of the reordered coefficients of predictor P$_3$ in (2) with $\ell = 3$ is the solution of the system of order conditions

$$\begin{bmatrix} 1 & 1 & 0 & 0 & 0 & 0 \\ c_2 & 0 & 1 & 0 & 0 & 0 \\ c_2^2/2! & 0 & 0 & 1 & 0 & 0 \\ c_2^3/3! & 0 & 0 & 0 & 1 & 0 \\ c_2^4/4! & 0 & 0 & 0 & 0 & 1 \\ c_2^5/5! & 0 & 0 & 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} a_{32} \\ a_{31} \\ \gamma_{32} \\ \gamma_{33} \\ \gamma_{34} \\ \gamma_{35} \end{bmatrix} = \begin{bmatrix} c_3 \\ c_3^2/2! \\ c_3^3/3! \\ c_3^4/4! \\ c_3^5/5! \\ c_3^6/6! \end{bmatrix}. \tag{22}$$

A truncated Taylor expansion of the right-hand side of (2) with $\ell = 3$ about $x_n$ gives

$$\sum_{j=0}^{13} S_3(j) h_{n+1}^j y_n^{(j)}$$

with coefficients

$$2S_3(j) = M^3(j,1:6)\mathbf{u}^3 = \frac{c_3^j}{j!}, \qquad j = 1,2,\ldots,6,$$

$$S_3(j) = a_{32} S_2(j-1), \qquad j = 7,8,\ldots,12.$$

### 4.4. PREDICTOR P$_4$

The 7-vector of the reordered coefficients of predictor P$_4$ in (2) with $\ell = 4$ is the solution of the system of order conditions

$$\begin{bmatrix} 1 & 1 & 1 & 0 & 0 & 0 & 0 \\ c_3 & c_2 & 0 & 1 & 0 & 0 & 0 \\ c_3^2/2! & c_2^2/2! & 0 & 0 & 1 & 0 & 0 \\ c_3^3/3! & c_2^3/3! & 0 & 0 & 0 & 1 & 0 \\ c_3^4/4! & c_2^4/4! & 0 & 0 & 0 & 0 & 1 \\ c_3^5/5! & c_2^5/5! & 0 & 0 & 0 & 0 & 0 \\ c_3^6/6! & c_2^6/6! & 0 & 0 & 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} a_{43} \\ a_{42} \\ a_{41} \\ \gamma_{42} \\ \gamma_{43} \\ \gamma_{44} \\ \gamma_{45} \end{bmatrix} \begin{bmatrix} c_4 \\ c_4^2/2! \\ c_4^3/3! \\ c_4^4/4! \\ c_4^5/5! \\ c_4^6/6! \\ c_4^7/7! \end{bmatrix}. \tag{23}$$

### 4.5. PREDICTOR P$_5$

The 7-vector of the reordered coefficients of predictor P$_5$ in (2) with $\ell = 5$ is the solution of the system of order conditions

$$\begin{bmatrix} 1 & 1 & 1 & 0 & 0 & 0 & 0 \\ c_4 & c_3 & 0 & 1 & 0 & 0 & 0 \\ c_4^2/2! & c_3^2/2! & 0 & 0 & 1 & 0 & 0 \\ c_4^3/3! & c_3^3/3! & 0 & 0 & 0 & 1 & 0 \\ c_4^4/4! & c_3^4/4! & 0 & 0 & 0 & 0 & 1 \\ c_4^5/5! & c_3^5/5! & 0 & 0 & 0 & 0 & 0 \\ c_4^6/6! & c_3^6/6! & 0 & 0 & 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} a_{54} \\ a_{53} \\ a_{51} \\ \gamma_{52} \\ \gamma_{53} \\ \gamma_{54} \\ \gamma_{55} \end{bmatrix} \begin{bmatrix} c_5 \\ c_5^2/2! \\ c_5^3/3! \\ c_5^4/4! \\ c_5^5/5! \\ c_5^6/6! \\ c_5^7/7! \end{bmatrix}. \tag{24}$$

## 4.6. THE COEFFICIENTS $a_{ij}$ OF $\mathrm{P}_i$, FOR $i = 6,7,8$ AND $j = 3,4,5$

It is numerically convenient first to solve for $a_{87}$ and $a_{76}$ from (11) and (12), and $a_{86}$ from (13). Next, we solve for the nine coefficients $a_{63}, a_{64}, a_{65}, a_{73}, a_{74}, a_{75}, a_{83}, a_{84}, a_{85}$ of predictors $\mathrm{P}_6$ to $\mathrm{P}_8$ simultaneously before solving for their other coefficients. These nine coefficients are solutions of the system of order conditions

$$\begin{bmatrix} c_5^5/5! & c_4^5/5! & c_3^5/5! & 0 & 0 & 0 & 0 & 0 & 0 \\ c_5^6/6! & c_4^6/6! & c_3^6/6! & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & c_5^5/5! & c_4^5/5! & c_3^5/5! & 0 & 0 & 0 \\ 0 & 0 & 0 & c_5^6/6! & c_4^6/6! & c_3^6/6! & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & c_5^5/5! & c_4^5/5! & c_3^5/5! \\ 0 & 0 & 0 & 0 & 0 & 0 & c_5^6/6! & c_4^6/6! & c_3^6/6! \\ 0 & 0 & b_6(1-c_6) & 0 & 0 & b_7(1-c_7) & 0 & 0 & b_8(1-c_8) \\ b_6(1-c_6)a_{53} & b_6(1-c_6)a_{43} & b_8(1-c_8)a_{86}+b_7(1-c_7)a_{76} & b_7(1-c_7)a_{53} & b_7(1-c_7)a_{43} & b_8(1-c_8)a_{87} & b_8(1-c_8)a_{53} & b_8(1-c_8)a_{43} & 0 \\ 0 & 0 & b_6(1-c_6)(c_8-c_6) & 0 & 0 & b_7(1-c_7)(c_8-c_7) & 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} a_{65} \\ a_{64} \\ a_{63} \\ a_{75} \\ a_{74} \\ a_{73} \\ a_{85} \\ a_{84} \\ a_{83} \end{bmatrix} = \mathbf{r} \tag{25}$$

where $\mathbf{r} = r(1:9)$ has components

$$
\begin{aligned}
r(1) &= c_6^6/6! \\
r(2) &= c_6^7/7! \\
r(3) &= c_7^6/6! - a_{76}c_6^5/5! \\
r(4) &= c_7^7/7! - a_{76}c_6^6/6! \\
r(5) &= c_8^6/6! - a_{87}c_7^5/5! - a_{86}c_6^5/5! \\
r(6) &= c_8^7/7! - a_{87}c_7^6/6! - a_{86}c_6^6/6! \\
r(7) &= -b_4(1-c_4)a_{43} - b_5(1-c_5)a_{53} \\
r(8) &= -b_5(1-c_5)a_{54}a_{43} \\
r(9) &= -b_4(1-c_4)(c_8-c_4)a_{43} - b_5(1-c_5)(c_8-c_5)a_{53}
\end{aligned}
$$

The equations for $r(7)$, $r(8)$ and $r(9)$ correspond to equations (15), (16) and (14), respectively.

## 4.7. PREDICTOR P$_6$

Since $a_{65}, a_{64}, a_{63}$ are already obtained from system (25), the remaining five unknown coefficients of predictor P$_6$ in (2) with $\ell = 6$ are in the following 5-vector of reordered coefficients,

$$\mathbf{u}^6 = [a_{61}, \gamma_{62}, \gamma_{63}, \gamma_{64}, \gamma_{65}]^T,$$

whose $i$ th component, $u_6(i)$, satisfies the order condition

$$u_6(i) = \frac{c_6^i}{i!} - \frac{1}{(i-1)!}\sum_{j=3}^{5} a_{6j}c_j^{i-1}, \qquad i = 1, 2, \ldots, 5. \tag{26}$$

## 4.8. PREDICTOR P$_7$

Since $a_{75}, a_{74}, a_{73}$ are already obtained from system (25), the remaining five unknown coefficients of predictor P$_7$ in (2) with $\ell = 7$ are in the following 5-vector of reordered coefficients,

$$\mathbf{u}^7 = [a_{71}, \gamma_{72}, \gamma_{73}, \gamma_{74}, \gamma_{75}]^T,$$

whose $i$ th component, $u_7(i)$, satisfies the order condition

$$u_7(i) = \frac{c_7^i}{i!} - \frac{1}{(i-1)!}\sum_{j=3}^{6} a_{7j}c_j^{i-1}, \qquad i = 1, 2, \ldots, 5, \tag{27}$$

where $a_{76}$ is obtained from (12).

### 4.9. PREDICTOR P$_8$

Since $a_{85}, a_{84}, a_{83}$ are already obtained from system (25), the remaining five unknown coefficients of predictor P$_8$ in (2) with $\ell = 8$ are in the following 5-vector of reordered coefficients,

$$\mathbf{u}^8 = [a_{81}, \gamma_{82}, \gamma_{83}, \gamma_{84}, \gamma_{85}]^T,$$

whose $i$ th component, $u_8(i)$, satisfies the order condition

$$u_8(i) = \frac{c_8^i}{i!} - \frac{1}{(i-1)!}\sum_{j=3}^{7} a_{8j} c_j^{i-1}, \qquad i = 1, 2, \ldots, 5, \tag{28}$$

where $a_{87}$ and $a_{86}$ are obtained from (11) and (13) respectively.

### 4.10. PREDICTOR P$_9$

The 11-vector of reordered coefficients of predictor P$_9$ in (2) with $\ell = 9$ is the solution of the system of order conditions

$$\begin{bmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 \\ c_8 & c_7 & c_6 & c_5 & c_4 & c_3 & 0 & 1 & 0 & 0 & 0 \\ c_8^2/2! & c_7^2/2! & c_6^2/2! & c_5^2/2! & c_4^2/2! & c_3^2/2! & 0 & 0 & 1 & 0 & 0 \\ c_8^3/3! & c_7^3/3! & c_6^3/3! & c_5^3/3! & c_4^3/3! & c_3^3/3! & 0 & 0 & 0 & 1 & 0 \\ c_8^4/4! & c_7^4/4! & c_6^4/4! & c_5^4/4! & c_4^4/4! & c_3^4/4! & 0 & 0 & 0 & 0 & 1 \\ b_9 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & b_9 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & b_9 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & b_9 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & b_9 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & b_9 & 0 & 0 & 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} a_{98} \\ a_{97} \\ a_{96} \\ a_{95} \\ a_{94} \\ a_{93} \\ a_{91} \\ \gamma_{92} \\ \gamma_{93} \\ \gamma_{94} \\ \gamma_{95} \end{bmatrix} = \mathbf{r}^9 \tag{29}$$

where $\mathbf{r}^9 = r_9(1:11)$ has components

$$\begin{aligned} r_9(i) &= c_9^i/i!, \qquad i = 1, 2, \ldots, 5 \\ r_9(6) &= b_8(1 - c_8) \\ r_9(7) &= b_7(1 - c_7) - b_8 a_{87} \\ r_9(8) &= b_6(1 - c_6) - (b_8 a_{86} + b_7 a_{76}) \\ r_9(9) &= b_5(1 - c_5) - (b_8 a_{85} + b_7 a_{75} + b_6 a_{65}) \\ r_9(10) &= b_4(1 - c_4) - (b_8 a_{84} + b_7 a_{74} + b_6 a_{64} + b_5 a_{54}) \\ r_9(11) &= b_3(1 - c_3) - (b_8 a_{83} + b_7 a_{73} + b_6 a_{63} + b_5 a_{53} + b_4 a_{43}) \end{aligned}$$

## 5. REGION OF ABSOLUTE STABILITY

To obtain the region of absolute stability, $R$, of HBT(11)9, we apply the predictors $\mathrm{P}_2$, $\mathrm{P}_3$, ..., $\mathrm{P}_9$ and the integration formula IF with constant step $h$ to the linear test equation

$$y' = \lambda y, \qquad y_0 = 1.$$

Thus we obtain

$$y_{n+c_\ell} = y_n + \lambda h_{n+1} \sum_{j=1}^{\ell-1} a_{\ell j} y_{n+c_j} + \sum_{j=2}^{3} (\lambda h_{n+1})^j \gamma_{\ell j} y_n, \qquad \ell = 2,3,\ldots,9, \tag{30}$$

and

$$y_{n+1} = y_n + \lambda h_{n+1} \sum_{j=1}^{9} b_j y_{n+c_j} + \sum_{j=2}^{5} (\lambda h_{n+1})^j \gamma_{1j} y_n. \tag{31}$$

If we replace $y_{n+c_\ell}$, for $\ell = 2,3,\ldots,9$, in (30)-(31) with the corresponding right-hand sides of (30), then (31) reduces to the following first-order difference equation and corresponding linear characteristic equation:

$$-r_s y_n + y_{n+1} = 0, \qquad -r_s + r = 0,$$

respectively. The root, $r_s$, of the characteristic equation is

$$r_s = 1 + \sum_{j=1}^{13} s_j \lambda^j h^j, \tag{32}$$

with coefficients

$$s_1 = 1.0$$
$$s_3 = 1.66666666666666\mathrm{e}-01$$
$$s_5 = 8.33333333333332\mathrm{e}-03$$
$$s_7 = 1.984126984126 97\mathrm{e}-04$$
$$s_9 = 2.75573192239858\mathrm{e}-06$$
$$s_{11} = 2.505210838526852\mathrm{e}-08$$
$$s_{13} = 4.110139481621576\mathrm{e}-10$$

$$s_2 = 5.00000000000000\mathrm{e}-01$$
$$s_4 = 4.16666666666666\mathrm{e}-02$$
$$s_6 = 1.38888888888888\mathrm{e}-03$$
$$s_8 = 2.48015873015873\mathrm{e}-05$$
$$s_{10} = 2.75573192239858\mathrm{e}-07$$
$$s_{12} = 9.751624424043938\mathrm{e}-10$$

A complex number $\lambda h$ is in $R$ if $r_s$ satisfies the root condition: $|r_s| \leq 1$ (see [8]).

The root condition is used to find the region of absolute stability of HBT(11)9 shown in grey in Fig. 1, with interval of absolute stability $(\alpha, 0) = (-5.40, 0)$. It is seen that HBT(11)9 has a larger scaled interval of absolute stability than DP(8,7)13M, namely, $5.4/13 = 0.41538 > 0.3938 = 5.12/13$.

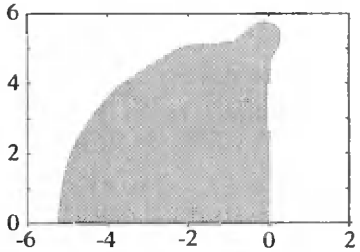


Fig. 1. Region of absolute stability of HBT(11)9.

## 6. CONTROLLING STEPSIZE

Generally, for a given tolerance TOL, the step size $h_{n+1}$ of a Taylor method of order $p$ is chosen as (see [11], [3])

$$h_{n+1} = k(\mathrm{TOL}, p) \| y_n^{(p)} / p! \|_\infty^{-1/p}, \tag{33}$$

where the function $k(\mathrm{TOL}, p)$ satisfies the equation $k^{p+1} / (1-k) = \mathrm{TOL}$ and the uniform norm of the vector $y$, $| y |_\infty$, is the largest component of $y$ in absolute value.

Since HBT(11)9 does not use derivatives of order higher than five, to determine the stepsize, for simplicity, we consider the following formula:

$$h_{n+1} = 1.6k(\mathrm{TOL}, 7)\left[\frac{\| y_n^{(3)} \|_\infty /3!}{\left[\| y_n^{(5)} \|_\infty /5!\right]^2}\right]^{1/7} \tag{34}$$

similar to an error estimator formula found in [8] or a procedure which is common for quadrature formulae [16] and [4].

## 7. NUMERICAL RESULTS

The derivatives, $y^{(2)}$ to $y^{(5)}$, are calculated at each integration step by known recurrence formulae (see, for example, [8], [11]).

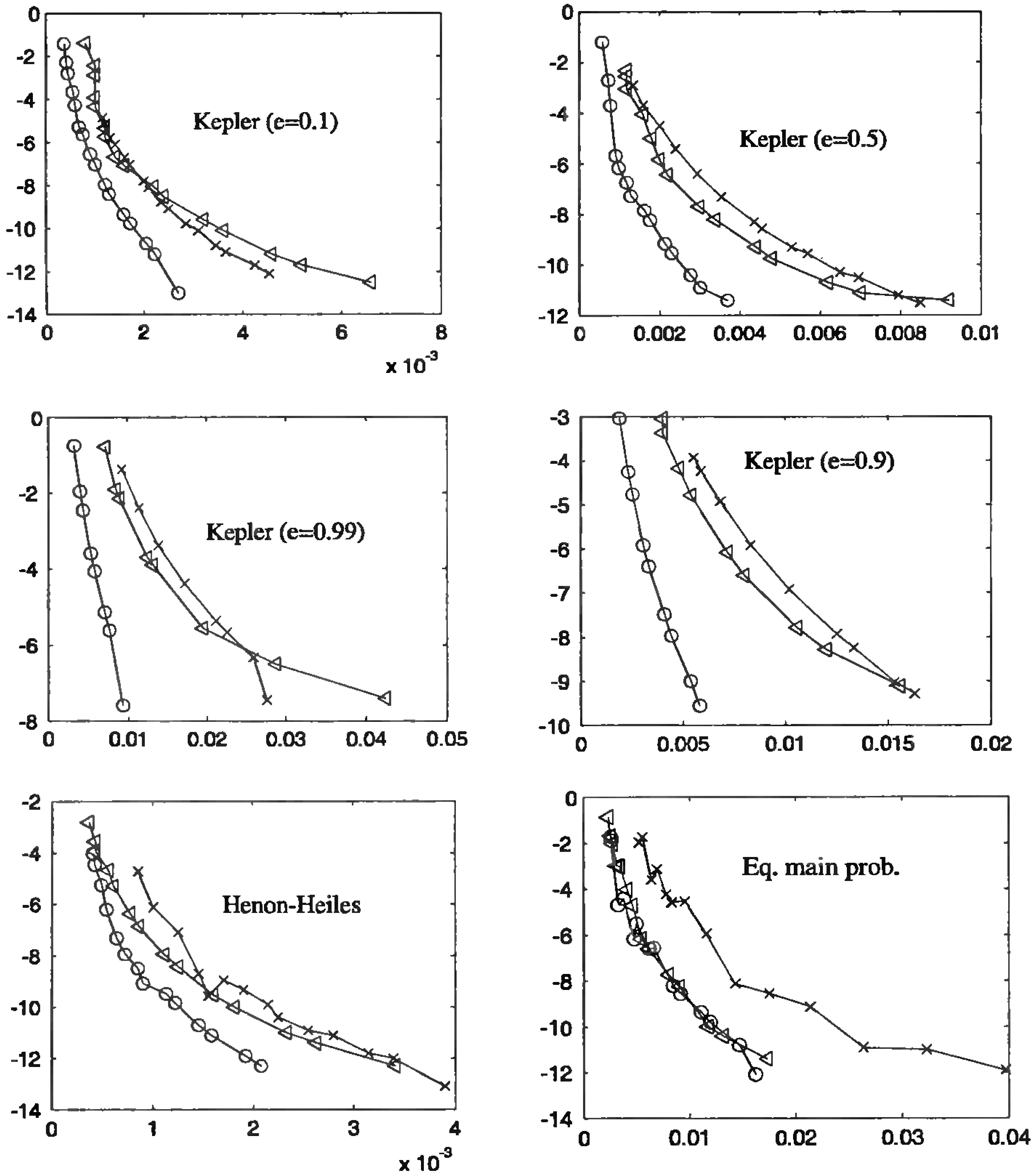
Computations were performed in C++ on a Mac with a dual 2.5 GHz PowerPC G5 and 4 GB DDR SSRAM running under Mac OS X Version 10.4.8.

### 7.1. COMPARISON BASED ON CPU

In our CPU comparison, HBT(11)9, T11, and DP(8,7) were applied to the following set of problems:

- Kepler's problem with eccentricity $\mathrm{e} = 0.1, 0.3, 0.5, 0.7, 0.9$ (DETEST D1-D5) and $\mathrm{e} = 0.99$,
- Hénon-Heiles' problem,
- the equatorial main problem.

The **maximum global error** (MGE) is taken to be $\max_n\{\| y_{n+1} - y(t_{n+1}) \|_\infty\}$ of the difference between the numerical and the analytic solutions at every integration step. In Fig. 2, CPU (horizontal axis) is plotted versus $\log_{10}(|\text{MGE}|)$ (vertical axis) for the above problems.



HBT(11)9 ○, T11 ×, DP(8,7)13M ▷

Fig. 2. CPU (horizontal axis) versus $\log_{10}(|\text{MGE}|)$ (vertical axis) for the listed problems .

The **CPU percentage efficiency gain** (CPU PEG) is defined by formula (cf. Sharp [19]),

$$(\text{CPU PEG})_i = 100\left[\frac{\sum_j \text{CPU}_{2,ij}}{\sum_j \text{CPU}_{1,ij}} - 1\right] \tag{35}$$

where $\text{CPU}_{1,ij}$ and $\text{CPU}_{2,ij}$ are the CPU of methods 1 and 2, respectively, associated with problem $i$, and $j = -\log_{10}(|\text{MGE}|)$. The CPU time was obtained from the curves which fit, in a least-squares sense, the data $(\log_{10}(|\text{MGE}|), \log_{10}(\text{CPU}))$ by means of Matlab's `polyfit`. The CPU PEG of HBT(11)9 over DP(8,7)13M and T11 for the above problems are listed in the middle part of Table 1.
It is seen from Fig. 2 and Table 1 that, at stringent tolerance, HBT(11)9 compares favorably with both DP(8,7)13M and T11 on the basis of CPU versus MGE and CPU PEG.

### 7.2. COMPARISON BASED ON THE NUMBER OF STEPS

The numerical performance of HBT(11)9 and T11 is also compared on the basis of the number of steps (NS), on Kepler's, Hénon-Heiles' and the equatorial main problems.

The **maximum global energy error** (MGEE) was obtained from the maximum of the absolute value of the relative error $H / H_0 - 1$ at every integration step where $H$ and $H_0$ are the values of the Hamiltonian at $t_{n+1}$ and at $t_0$, respectively.

Table 1. CPU and NS PEG of HBT(11)9 over DP(8,7)13M and T11 for the listed problems.

| Problem | CPU PEG of HBT(11)9 over: | | NS PEG of HBT(11)9 over: | |
|---|---|---|---|---|
| | DP(8,7)13M | T11 | DP(8,7)13M | T11 |
| Kepler (e = 0.1) | 95% | 75% | 93% | 34% |
| Kepler (e = 0.3) | 117% | 158% | 109% | 105% |
| Kepler (e = 0.5) | 112% | 148% | 104% | 108% |
| Kepler (e = 0.7) | 146% | 203% | 122% | 135% |
| Kepler (e = 0.9) | 148% | 178% | 131% | 122% |
| Kepler (e = 0.99) | 204% | 193% | 126% | 124% |
| Hénon-Heiles | 48% | 76% | 89% | 42% |
| Eq. main prob. | 5% | 108% | 21% | 46% |
| B1 | 45% | 47% | | |
| B5 | 86% | 201% | | |
| E2 | 30% | 66% | | |
| Arenstorf | 103% | 241% | | |

The Hamiltonians of Kepler's, Hénon-Heiles' and the equatorial main problems [3] are

$$H_{\text{Kepler}} = \frac{1}{2}\left(y_3^2 + y_4^2\right) - 1/\sqrt{y_1^2 + y_2^2} \tag{36}$$

$$H_{\text{Henon-Heiles}} = \frac{1}{2}\left(X^2 + Y^2\right) + \frac{1}{2}\left(x^2 + y^2\right) + \varepsilon y\left(x^2 - \frac{1}{3}y^2\right) \tag{37}$$

$$H_{\text{eq. main prob.}} = \frac{1}{2}\left(P^2 + \frac{\Lambda^2}{\rho^2} + Z^2\right) + \frac{\mu}{r} + \frac{\alpha^2 J_2 \mu P_2(u)}{r^3} \tag{38}$$

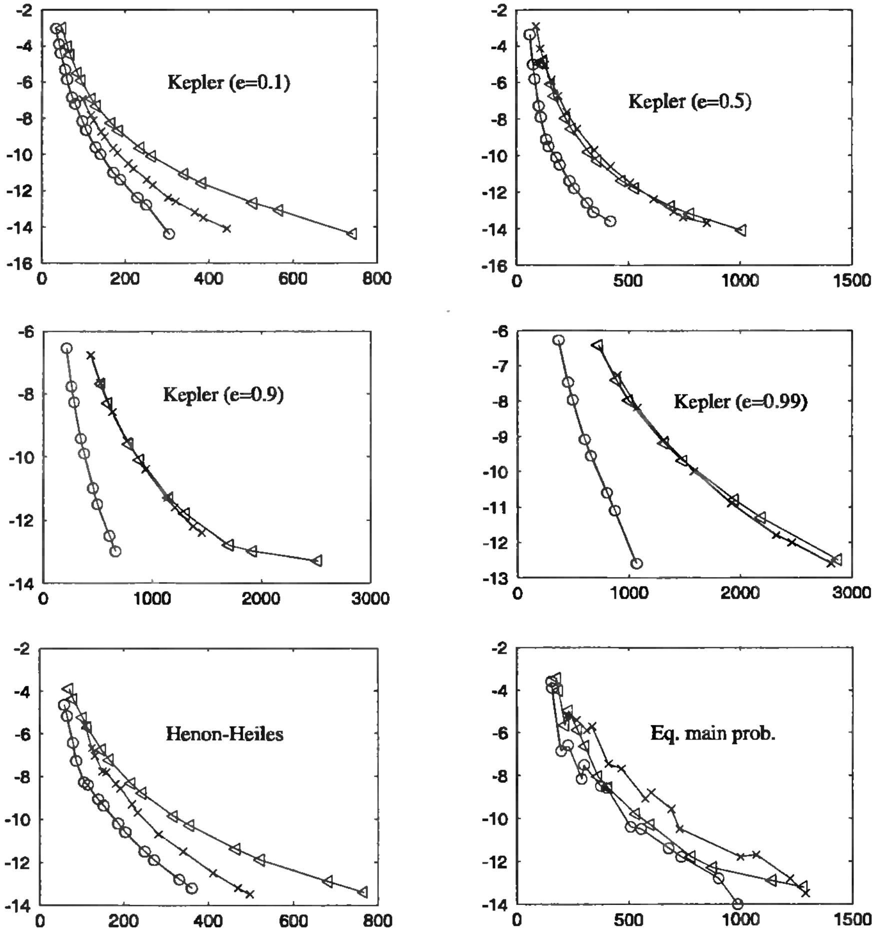
respectively, where, in (38), $u = z / r$, $r = \sqrt{\rho^2 + z^2}$ and $P_2(x) = (3x^2 - 1)/2$ is the Legendre polynomial of degree 2.

In Fig. 3, the number of step (horizontal axis) is plotted versus $\log_{10}(|\text{MGEE}|)$ (vertical axis) for the problems on hand.

The **number of step percentage efficiency gain** $(\text{NS PEG})_i$ for the $i$ th problem is defined by the formula

$$(\text{NS PEG})_i = 100\left[\frac{\sum_j \text{NS}_{\text{T},ij}}{\sum_j \text{NS}_{\text{HBT},ij}} - 1\right], \tag{39}$$

where $\text{NS}_{\text{T},ij}$ and $\text{NS}_{\text{HBT},ij}$ are the number of steps used by methods T11 and HBT(11)9, respectively, to integrate from $t_0$ to $t_f$, and $j = -\log_{10}(|\text{MGEE}|)$. The number of steps (NS) was obtained from the curves which fit the data $(\log_{10}(|\text{MGEE}|), \log_{10}(\text{NS}))$ in the least squares sense.



HBT(11)9 ○, T11 ×, DP(8,7)13M ▷

Fig. 3. Number of steps (horizontal axis) versus $\log_{10}(|\text{MGEE}|)$ (vertical axis) for the problems in hand.

It is observed that HBT(11)9 performs better than T11 on the basis of the number of steps versus MGEE shown in Fig. 3 and the number of step percentage efficiency gain listed in the rightmost part of Table 1.

It is to be noted that HBT(11)9 uses five derivatives of $y$ compared to eleven for T11. The numerical results show that a combination of high-order derivatives with a Runge-Kutta method achieves a high degree of accuracy.

## 8. CONCLUSION

A one-step 9-stage Hermite-Birkhoff-Taylor method of order 11, called HBT(11)9, was constructed by solving Vandermonde-type systems satisfying Taylor- and Runge-Kutta-type order conditions. By construction, HBT(11)9 uses lower order derivatives than the traditional Taylor method of order 11. The stability region of HBT(11)9 has a remarkably good shape. The stepsize is controlled by a formula which uses $y_n^{(3)}$ and $y_n^{(5)}$. On the basis of CPU versus the maximum global error, and the number of steps versus the maximum global energy error, HBT(11)9 wins over both DP(8,7)13M and T11 in solving the problems in hand.

## APPENDIX A. FORMULAE FOR HBT(11)9

The formulae for the one-step HBT(11)9 using the offstep points listed in (17) are:

$$\begin{aligned} y_{n+c_2} = y_n &+ 0.27921731436446077438372923301803\mathrm{e}+00\, h_{n+1} f_n + \\ &+ 0.38981154320451056548200709778768\mathrm{e}-01\, h_{n+1}^2 y_n^{(2)} + \\ &+ 0.36280710733943136559513194218467\mathrm{e}-02\, h_{n+1}^3 y_n^{(3)} + \\ &+ 0.25325506535913667870418670407485\mathrm{e}-03\, h_{n+1}^4 y_n^{(4)} + \\ &+ 0.14142639839754825204811385189360\mathrm{e}-04\, h_{n+1}^5 y_n^{(5)} \end{aligned}$$

$$\begin{aligned} y_{n+c_3} = y_n &+ 0.11734695722473867471661865760511\mathrm{e}+00\, h_{n+1} f_{n+c_2} + \\ &+ 0.20840657620046556206439878091593\mathrm{e}+00\, h_{n+1} f_n + \\ &+ 0.20292380024370036960387719454519\mathrm{e}-01\, h_{n+1}^2 y_n^{(2)} + \\ &+ 0.11869226429345444634690707303528\mathrm{e}-02\, h_{n+1}^3 y_n^{(3)} + \\ &+ 0.43443173577338189126711268038656\mathrm{e}-04\, h_{n+1}^4 y_n^{(4)} + \\ &+ 0.84910603776134311848874368025906\mathrm{e}-06\, h_{n+1}^5 y_n^{(5)} \end{aligned}$$

$$\begin{aligned} y_{n+c_4} = y_n &+ 0.13690478342886178716938843387260\mathrm{e}+00\, h_{n+1} f_{n+c_3} + \\ &+ 0.0\mathrm{e}+00\, h_{n+1} f_{n+c_2} \\ &+ 0.24314100556720982240846524440193\mathrm{e}+00\, h_{n+1} f_n + \\ &+ 0.27620183922059216973861062590877\mathrm{e}-01\, h_{n+1}^2 y_n^{(2)} + \\ &+ 0.18847891968821701433791262060701\mathrm{e}-02\, h_{n+1}^3 y_n^{(3)} + \\ &+ 0.80483842406781629701569255062404\mathrm{e}-04\, h_{n+1}^4 y_n^{(4)} + \\ &+ 0.18352527233352486872994232297006\mathrm{e}-05\, h_{n+1}^5 y_n^{(5)} \end{aligned}$$

$$
\begin{aligned}
y_{n+c_5} = y_n &+ 4.89327430135532609839861592043 17\mathrm{e}+00 h_{n+1} f_{n+c_4} + \\
&- 7.84941424114178659754214096519 66\mathrm{e}+00 h_{n+1} f_{n+c_3} + \\
&+ 0.0\mathrm{e}+00 h_{n+1} f_{n+c_2} \\
&+ 3.58185577957003797571736429365 06\mathrm{e}+00 h_{n+1} f_n + \\
&+ 0.89306628781530294412722895316551\mathrm{e}+00 h_{n+1}^2 y_n^{(2)} + \\
&+ 0.10392225185475299711970223172354\mathrm{e}+00 h_{n+1}^3 y_n^{(3)} + \\
&+ 0.68425817883121991387825431868206\mathrm{e}-02 h_{n+1}^4 y_n^{(4)} + \\
&+ 0.22876885553059258475652435914118\mathrm{e}-03 h_{n+1}^5 y_n^{(5)}
\end{aligned}
$$

$$
\begin{aligned}
y_{n+c_6} = y_n &+ 0.18705263040983843260252002847265\mathrm{e}+00 h_{n+1} f_{n+c_5} + \\
&- 2.0844005750645413591144920195002\mathrm{e}+00 h_{n+1} f_{n+c_4} + \\
&+ 4.2958342051718070421139069373008\mathrm{e}+00 h_{n+1} f_{n+c_3} + \\
&+ 0.0\mathrm{e}+00 h_{n+1} f_{n+c_2} + \\
&- 1.7137868058462236681641819548824\mathrm{e}+00 h_{n+1} h_{n+1} f_n + \\
&- 0.48985063231661957748720178069248\mathrm{e}+00 h_{n+1}^2 y_n^{(2)} + \\
&- 0.60515093159550711661779057964195\mathrm{e}-01 h_{n+1}^3 y_n^{(3)} + \\
&- 0.41595380328525402823589894252955\mathrm{e}-02 h_{n+1}^4 y_n^{(4)} + \\
&- 0.14437767105118407830394524352110\mathrm{e}-03 h_{n+1}^5 y_n^{(5)}
\end{aligned}
$$

$$
\begin{aligned}
y_{n+c_7} = y_n &+ 0.30989389672890554665807005124822\mathrm{e}+00 h_{n+1} f_{n+c_6} + \\
&+ 0.33647558747357717824967958660402\mathrm{e}+00 h_{n+1} f_{n+c_5} + \\
&- 26.478506375324671073355351111011\mathrm{e}+00 h_{n+1} f_{n+c_4} + \\
&+ 47.222749333325735677691659825296\mathrm{e}+00 f_{n+c_3} + \\
&+ 0.0\mathrm{e}+00 f_{n+c_2} + \\
&- 20.595232860565683176689473845122\mathrm{e}+00 h_{n+1} f_n + \\
&- 5.4263405539713950804028761913974\mathrm{e}+00 h_{n+1}^2 y_n^{(2)} + \\
&- 0.64796710884005649219800265659289\mathrm{e}+00 h_{n+1}^3 y_n^{(3)} + \\
&- 0.43461055304334359110125728903391\mathrm{e}-01 h_{n+1}^4 y_n^{(4)} + \\
&- 0.14747577522970799805523001545102\mathrm{e}-02 h_{n+1}^5 y_n^{(5)}
\end{aligned}
$$

$$
\begin{aligned}
y_{n+c_8} = y_n &+ 0.73545649421025738446857064434108\mathrm{e}+00 h_{n+1} f_{n+c_7} + \\
&- 2.8046503005389438917777274216335\mathrm{e}+00 h_{n+1} f_{n+c_6} + \\
&+ 2.4310902636662130713676488092095\mathrm{e}+00 h_{n+1} f_{n+c_5} + \\
&+ 47.005874548299616378597396578655\mathrm{e}+00 h_{n+1} f_{n+c_4} + \\
&- 85.530427688187376154049142427554\mathrm{e}+00 h_{n+1} f_{n+c_3} + \\
&+ 0.0\mathrm{e}+00 h_{n+1} f_{n+c_2} + \\
&+ 39.087312960190737609578582044537\mathrm{e}+00 h_{n+1} f_n + \\
&+ 10.239152737251216021340756008011\mathrm{e}+00 h_{n+1}^2 y_n^{(2)} + \\
&+ 1.2240487287625932711944296054422\mathrm{e}+00 h_{n+1}^3 y_n^{(3)} + \\
&+ 0.82287938650991655598093344878360\mathrm{e}-01 h_{n+1}^4 y_n^{(4)} + \\
&+ 0.27963172725310360887791680345844\mathrm{e}-02 h_{n+1}^5 y_n^{(5)}
\end{aligned}
$$

$$\begin{aligned}
y_{n+c_9} = y_n &+ 0.38868049144002532168874211206637\mathrm{e}+00 h_{n+1} f_{n+c_8} + \\
&- 2.5222058120884045886998467007211\mathrm{e}+00 h_{n+1} f_{n+c_7} + \\
&+ 12.542349667143804316191097678012\mathrm{e}+00 h_{n+1} f_{n+c_6} + \\
&- 11.085862883196452045486175542222\mathrm{e}+00 h_{n+1} f_{n+c_5} + \\
&- 116.79158724091865449522082531143\mathrm{e}+00 h_{n+1} f_{n+c_4} + \\
&+ 220.44468870550405129888355543344\mathrm{e}+00 h_{n+1} f_{n+c_3} + \\
&+ 0.0\mathrm{e}+00 h_{n+1} f_{n+c_2} + \\
&- 101.97606292788436980735654766914\mathrm{e}+00 h_{n+1} f_n + \\
&- 26.928910176041650041777075194680\mathrm{e}+00 h_{n+1}^2 y_n^{(2)} + \\
&- 3.2334218544321017425241613689870\mathrm{e}+00 h_{n+1}^3 y_n^{(3)} + \\
&- 0.21794841028283032621984285869501\mathrm{e}+00 h_{n+1}^4 y_n^{(4)} + \\
&- 0.74111457430224934039870558289608\mathrm{e}-02 h_{n+1}^5 y_n^{(5)}
\end{aligned}$$

$$\begin{aligned}
y_{n+1} = y_n &+ 0.22310509503784392207559471537063\mathrm{e}-01 h_{n+1} f_{n+c_9} \\
&+ 0.11509465588695300265469403785365\mathrm{e}+00 h_{n+1} f_{n+c_8} \\
&+ 0.13867343057444646238526711766188\mathrm{e}+00 h_{n+1} f_{n+c_7} \\
&- 0.53588468963958452810456885263058\mathrm{e}-11 h_{n+1} f_{n+c_6} \\
&+ 0.21142885685125544800118700950983\mathrm{e}+00 h_{n+1} f_{n+c_5} \\
&+ 0.26963155018059286312026670753891\mathrm{e}+00 h_{n+1} f_{n+c_4} \\
&+ 0.0\mathrm{e}+00 h_{n+1} f_{n+c_3} + 0.0\mathrm{e}+00 h_{n+1} f_{n+c_2} \\
&+ 0.24286099700832667852742150117970\mathrm{e}+00 h_{n+1} f_n \\
&+ 0.26201759270767070765199711471931\mathrm{e}-01 h_{n+1}^2 y_n^{(2)} \\
&+ 0.15832358538292644689861943822241\mathrm{e}-02 h_{n+1}^3 y_n^{(3)} \\
&+ 0.54129494760573755982925911502439\mathrm{e}-04 h_{n+1}^4 y_n^{(4)} \\
&+ 0.84819956731509655706403853355909\mathrm{e}-06 h_{n+1}^5 y_n^{(5)}
\end{aligned}$$

## APPENDIX B. RECURRENT COMPUTATION OF HIGH-ORDER DERIVATIVES

To advance integration from $x_n$ to $x_{n+1}$, once $y_{n+1}$ is obtained by formula (3), the function $g$, with input $(x_{n+1}, y_{n+1})$,

$$[f_{n+1}, f_{n+1}^{(1)}, \dots, f_{n+1}^{(4)}] = g(x_{n+1}, y_{n+1})$$

outputs $f_{n+1}$ and $f_{n+1}^{(1)}$ to $f_{n+1}^{(4)}$ by means of the recurrent power series method. In adding, multiplying or taking powers of input power series, this method computes, in a recurrent way, the $k$ th term of the output power series as acombination of the preceding terms of the input series.

For precision and efficiency, Horner's scheme is used to evaluate the second summation in (2) and (3) in the form of nested polynomials in $h_{n+1}$.

## BIBLIOGRAPHY

[1] Arenstorf R.F., 1963: Periodic solutions of the restricted three-body problem representing analytic continuations of Keplerian elliptic motions, Amer. J. Math., vol. LXXXV, 27-35.

[2] Barrio R., 2006: Sensitivity analysis of ODEs/DAEs using the Taylor series method, SIAM J. Sc. Comp., vol. 27(6), 1929-1947.

[3] Barrio R., Blesa F., Lara M., 2005: VSVO formulation of the Taylor method for the numerical solution of ODEs, Comput. Math. Applic., vol. 50, 93-111.

[4] Berntsen J., Espelid T.O., 1991: Error estimation in automatic quadrature routines, ACM Trans. Math. Software, vol. 17, 233-255.

[5] Corliss G.F., Chang Y.F., 1982: Solving ordinary differential equations using Taylor series, ACM Trans. Math. Software, vol. 8(2), 114-144.

[6] Davis P.J., Rabinowitz P. 1967: Numerical Integration, Blaisdell, Waltham MA.

[7] Deprit A., Zahar R.M.W., 1966: Numerical integration of an orbit and its concomitant variations, Z. Angew. Math. Phys., vol. 17, 425-430.

[8] Hairer E., Nørsett S.P., Wanner G., 1993: Solving Ordinary Differential Equations I. Nonstiff Problems. Section III.8, Springer-Verlag, Berlin.

[9] Hoefkens J., Berz M., Makino K., 2003: Computing validated solutions of implicit differential equations, Adv. Comput. Math., vol. 19, 231-253.

[10] Hull T.E., Enright W.H., Fellen B.M. and Sedgwick, A. E.: Comparing numerical methods for ordinary differential equations, SIAM J. Numer. Anal., vol. 9 (1972), 603-637.

[11] Lara M., Elipe A., Palacios M., 1999: Automatic programming of recurrent power series, Math. Comput. Simul., vol. 49, 351-362.

[12] Nedialkov N.S., Jackson K.R., Corliss G.F., 1999: Validated solutions of initial value problems for ordinary differential equations, Appl. Math. Comput., vol. 105, 21-68.

[13] Nguyen-Ba T., Kengne E., Vaillancourt R.: One-step 4-stage Hermite-Birkhoff-Taylor ODE Solver of order 12, submitted to Can. Appl. Math. Quarterly.

[14] Nguyen-Ba T., Bozic V., Kengne E., Vaillancourt R., 2008: One-step 7-stage Hermite-Birkhoff-Taylor ODE Solver of order 13, International J. Pure Appl. Math., vol. 43(4), 569-592.

[15] Nguyen-Ba T., Bozic V., Kengne E., Vaillancourt R., 2007: One-step 4-stage Hermite-Birkhoff-Taylor ODE Solver of order 14, Scientific Proceedings of Riga Technical University, vol. 33, 49th issue, 6-25.

[16] Piessens R., Doncker-Kapenga E., de Überhuber C.W., Kahaner D.K., 1983: QUADPACK. A subroutine package for automatic integration, Springer Series in Comput. Math., vol. 1.

[17] Prince P.J., Dormand J.R., 1981: High order embedded Runge-Kutta formulae, J. Comput. Appl. Math., vol. 7(1), 67-75.

[18] Rabe E., 1961: Determination and survey of periodic Trojan orbits in the restricted problem of three bodies, Astronomical J., vol. 66(9), 500-513.

[19] Sharp P.W., 1991: Numerical comparison of explicit Runge-Kutta pairs of orders four through eight, ACM Trans. Math. Software, vol. 17, 387-409.

[20] Steffensen J.F., 1956: On the restricted problem of three bodies, Danske Vid. Selsk., Mat.-fys. Medd., vol. 30(18). 17 p.

# JEDNOKROKOWA DZIEWIĘCIOETAPOWA METODA HERMITA-BIRKHOFFA-TAYLORA DLA ROZWIĄZYWANIA RÓWNAŃ RÓŻNICZKOWYCH ZWYCZAJNYCH RZĘDU JEDENASTEGO

## Streszczenie

Jednokrokowa dziewięcioetapowa metoda Hermita-Birkhoffa-Taylora rzędu 11, oznaczona HBT(11)9, służy do rozwiązywania niesztywnych układów równań różniczkowych pierwszego rzędu mających formę $y' = f(x, y)$, $y(x_0) = y_0$. W metodzie tej wykorzystuje się zarówno $y'$ jak i wyższe pochodne, od $y^{(2)}$ do $y^{(5)}$, tak jak w metodach Taylora; jest ona powiązana z metodą Rungego-Kutty 9-tego rzędu. Dobierając współczynniki rozwinięcia Taylora rozwiązania numerycznego przez porównanie z rzeczywistym rozwiązaniem, otrzymuje się zbiór warunków niezbędnych do osiągnięcia zadanego rzędu. Powyższe warunki, zapisane w formie układu równań liniowych typu Vandermonda, dają w rozwiązaniu współczynniki metody. Nowa metoda posiada szerszy przedział bezwzględnej stabilności niż metoda Dormanda-Prince'a DP(8,7)13M. Wielkość kroku jest kontrolowana za pomocą pochodnych $y^{(3)}$ i $y^{(5)}$. Metoda HBT(11)9 jest lepsza – rozpatrując liczbę kroków, czas użycia procesora i maksymalny błąd globalny – od metody Dormanda-Prince'a DP(8,7)13M oraz metody Taylora rzędu 11, które są często używane w testowaniu systemów służących do rozwiązywania równań różniczkowych zwyczajnych wyższego rzędu. Przedstawione rezultaty numeryczne pokazują zalety dodania wyższych pochodnych do metody Rungego-Kutty.

Słowa kluczowe: ogólna metoda liniowa dla rozwiązywania niesztywnych równań różniczkowych zwyczajnych, metoda Hermita-Birkhoffa-Taylora, maksymalny błąd globalny, liczba ewaluacji funkcji, czas pracy procesora.

UNIWERSYTET TECHNOLOGICZNO-PRZYRODNICZY
IM. JANA I JĘDRZEJA ŚNIADECKICH W BYDGOSZCZY
ZESZYTY NAUKOWE NR 250
TELEKOMUNIKACJA I ELEKTRONIKA 11 (2008), 53-63

# PROSPECTS FOR INTERSTITIAL HYPERTHERMIA OF VERY SMALL TUMORS WITH 9-30 GHz MICROWAVES

Victor L. Granatstein*, John C. Rodgers, Binyam Yeshitla, Sri Surapaneni

Electrical and Computer Engineering Department
Institute for Research in Electronics and Applied Physics
University of Maryland
College Park, MD 20742-3511, USA
*vlg@umd.edu

*Summary*: The minimum achievable volume for local deposition of microwave energy in biological material has been assessed in the frequency range 9 GHz to 30 GHz. Both numerical simulation of deposition in human tissue and experiments with chicken breast tissue (similar to human muscle) were employed. With a very short monopole antenna at the end of a miniature co-axial cable and a 10 Watt, 1 second pulse of 9 GHz microwaves, a sphere with diameter of 1.5 millimeters was preferentially heated to temperatures required for cell destruction. Simulations indicate that if frequency were raised to 30 GHz the diameter of the heated sphere would shrink to 1 millimeter. Tumors in humans are often first detected when they are about 1 millimeter in diameter. The results reported are preliminary in nature being obtained as part of a training project for M.S. students.

Keywords: microwave hyperthermia, interstitial applicator, small tumor destruction, X-band, K-band.

## 1. INTRODUCTION

Tumors are often first detected when they are very small, about 1 mm in diameter. Destroying tumors by microwave hyperthermia ideally requires that the microwave energy be deposited in the tumor volume with minimal heating of surrounding healthy tissue; the situation becomes more challenging if the tumor is located deep inside the body (perhaps many centimeters from the body surface). Previous localized microwave hyperthermia studies have not been aimed at such small tumors. Arrays of microwave aperture antennas outside the body have been used to achieve localized microwave energy concentration by phase interference [1], but this technique inevitably involves some undesirable heating of healthy tissue located between the tumor site and the body surface. Alternatively, microwave energy has been injected by inserting interstitially a coaxial cable terminated by an antenna which is placed in close proximity to the tumor site; the coaxial cable is either inserted through a natural body orifice or is part of a hypodermic needle assembly which pierces the skin. The microwave frequency used in such interstitial procedures has been in the range 433 MHz to 2450 MHz and the heated region has ranged in linear extent from 7 cm to 2.5 cm respectively [2]. Much greater localization of hyperthermia therapy can be obtained with a laser fiberoptic interstitial delivery system [3], but the profile of energy deposition with microwaves is very different than with lasers, and may be advantageous if comparable localization were achievable.

We have explored the possibility of realizing strongly confined deposition of microwave energy in biological material in a volume of millimeter extent. The microwave frequency has been increased to the range 9 GHz to 30 GHz beyond the usual range for interstitial hyperthermia. There have been a number of previous studies of the effect of varying microwave frequency in hyperthermia [4] but these have generally considered frequencies lower than those in the current study.

In addition to raising frequency, the following two phenomena have been exploited to realize strong localization:

1. Local hot spot formation in a small volume in the near field of a monopole antenna formed by the inner conductor of a co-axial cable extending a small fraction of a wavelength past the end of the cable [5] and [6] and rapid spatial decay of the temperature profile [7].
2. The well-known phenomenon of decreasing penetration depth of microwaves into biological material as the microwave frequency increases [8]]. Specifically the microwave frequency spans the range where the depth of penetration of the microwaves into some types of biological tissue goes through 1 millimeter.

The deposition profile of microwave energy in biological tissue depends on microwave frequency, f, and the values of relative permittivity, $\varepsilon_r$ , and conductivity, σ, of the tissue both of which are frequency dependent. Values of $\varepsilon_r$ and σ [9] are displayed in Table 1 for a number of frequency values and for two types of human tissue, muscle and breast fat. Also displayed in Table 1 are the derived quantities loss tangent, tan δ, absorption length, $L_A$, and wavelength in the tissue, λ, where

$$\tan\delta = 18\,\sigma / (f_{GHz}\,\varepsilon_r), \tag{1}$$

$$L_A(mm) = \frac{1}{0.0148 f_{GHz}\sqrt{\varepsilon_r[\sqrt{1+(\tan\delta)^2}-1]}}, \tag{2}$$

and

$$\lambda\,(\mathrm{mm}) = 300 / [\,\varepsilon_r^{1/2}\,f_{GHz}]\,. \tag{3}$$

Table 1. Microwave properties of human tissue.

| Tissue Name | Frequency $f_{GHz}$ [GHz] | Conductivity σ [S/m] | Relative Permittivity $\varepsilon_r$ | Loss Tangent tan δ | Absorption Length $L_A$ [mm] | Wavelength in Tissue λ [mm] |
|---|---|---|---|---|---|---|
| Muscle | 9 | 9.192 | 44.126 | 0.4161 | 3.92 | 5.018 |
| Muscle | 16 | 19.236 | 35.168 | 0.6171 | 1.70 | 3.16 |
| Muscle | 23 | 28.303 | 28.224 | 0.7838 | 1.06 | 2.455 |
| Muscle | 30 | 35.487 | 23.157 | 0.9182 | 0.782 | 2.07 |
| Breast Fat | 9 | 0.675 | 4.003 | 0.3366 | 15.96 | 16.66 |
| Breast Fat | 30 | 1.406 | 2.913 | 0.2892 | 6.51 | 5.85 |
| Breast Fat | 100 | 1.836 | 2.589 | 0.1275 | 4.66 | 1.86 |

As may be seen in the table, in the frequency range $9 < f_{GHz} < 30$, the absorption length in muscle decreases to sub-millimeter values as frequency is raised; also, for $f_{GHz} = 23$ and for $f_{GHz} = 30$, the absorption length is smaller than a half wavelength. This is not true for beast fat.

## 2. SIMULATIONS

The patterns of microwave power deposition in theses two types of tissue has been simulated using the HFSS Code [10]. HFSS is a finite element, 3-dimemnsional electromagnetic field simulator. The geometry being simulated is shown in Fig. 1.

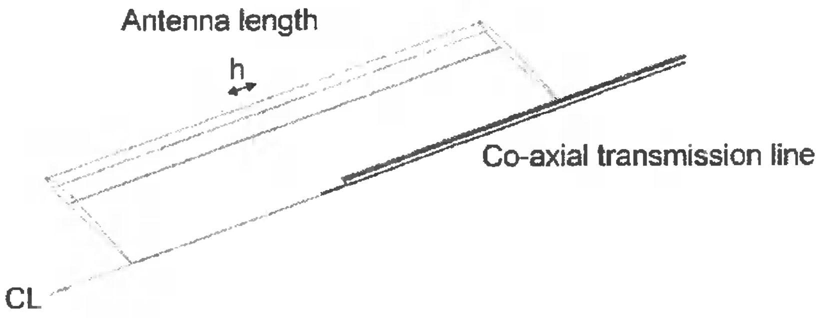


Fig. 1. Geometry of coaxial line terminated by monopole antenna inside biological tissue.

In Fig. 1, azimuthal symmetry is assumed around a center line (CL). The geometry consists of a co-axial transmission line terminated in a monopole antenna; both the transmission line and the antenna are surrounded by biological tissue. The dimensions of the co-axial cable are as follows: inner conductor o.d. = 0.020 mm; dielectric o.d. = 0.66 mm; outer conductor o.d. = 0.86 mm. The relative permittivity of the dielectric is 2.01.

The conditions used in various simulations are the following:

- antenna length, h (0.9 mm, 2.5 mm),
- microwave frequency, f (9 GHz, 16 GHz, 23 GHz, 30 GHz),
- biological tissue (human muscle, human breast fat).

The results of the simulations are color coded plots of the microwave power loss density (Watts /meter$^3$). The color coding is indicated in Fig. 2.

In Fig. 3a, the spatial pattern of power loss density at f = 9 GHz is displayed for the two cases of breast fat tissue and muscle tissue with antenna length h = 2.5 mm. These patterns should be compared with the patterns in Fig. 3b where the antenna length is 0.9 mm. With h = 2.5 mm, it is seen that the region of most intense power loss density (red) has a maximum diameter of ~2 millimeter but is elongated in shape with a length in breast fat of ~6 mm and a length in muscle of ~5 mm. When h = 0.9 mm, the patterns are dramatically shortened and more spherical in shape; the maximum diameter is still about ~2 mm, but the length is now only 2.6 mm. The pattern produced by the shorter antenna length would be more suitable for hyperthermia of a very small tumor with minimal heating of surrounding healthy tissue.

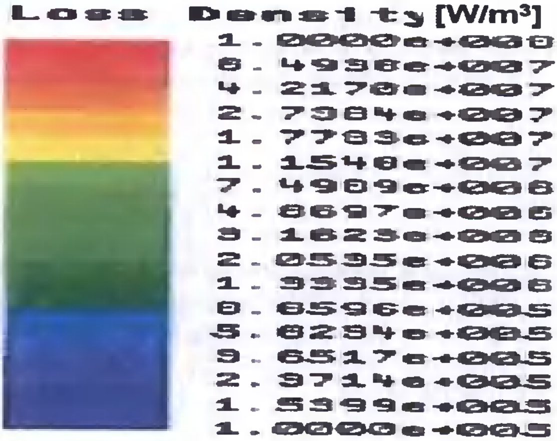


Fig. 2. Color coding for relative power loss density.

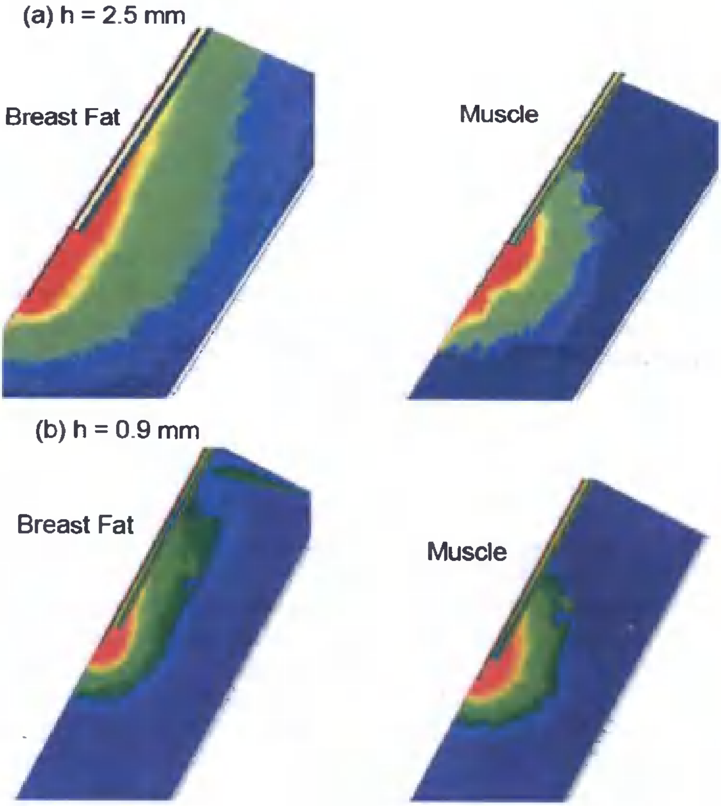


Fig. 3. Power loss density patterns for f = 9 GHz: (a) antenna length h = 2.5 mm; (b) antenna length h = 0.9 mm.

In Fig. 4, the power density absorption pattern in human muscle for an antenna length of 0.9 mm is compared at four values of microwave frequency; viz. 9 GHz, 16 GHz, 23 GHz and 30 GHz. It is seen that the dimensions of the red region of highest power loss density decreases monotonically with increasing frequency. At 9 GHz the length of the red region is 2.6 mm and its maximum radius is 1.1 mm. At 30 GHz, the length of the red region has decreased to 1.8 mm and its maximum radius is 0.7 mm. Note from Table 1, that as frequency rises from 9 GHz to 30 GHz, the microwave power absorption length in human muscle decreases from 3.92 mm to 0.782 mm comparable to the linear extent of the red regions in Fig. 4.

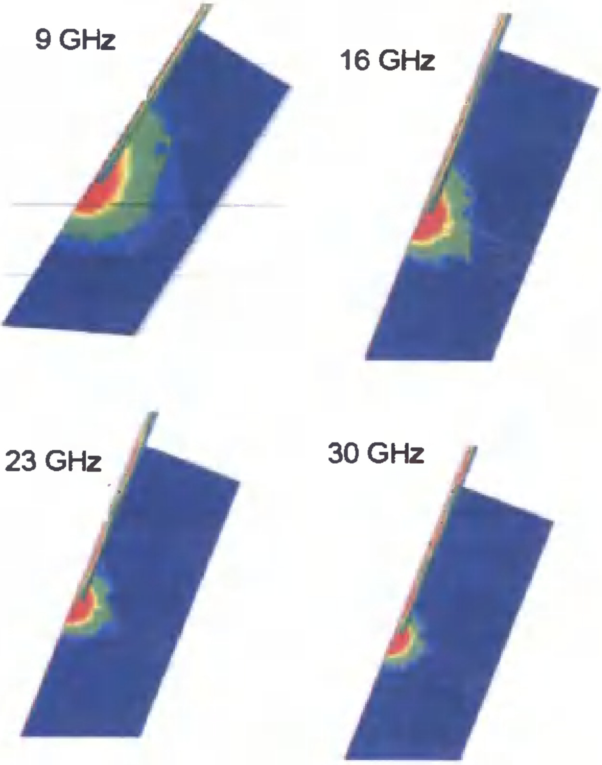


Fig. 4. Power loss density patterns in human muscle tissue at various frequencies (antenna length h = 0.9 mm).

In Fig. 5, the study of the effect of microwave frequency on the power loss density pattern is repeated for the case of human breast fat tissue. In this case, no decrease in the size of the absorption region is observed as frequency rises. Note from Table 1, as frequency rises from 9 GHz to 30 GHz, the microwave absorption length in human breast fat decreases from 15.96 mm to 6.51 mm. These absorption length values are considerably larger than the linear extent of the red regions in Fig. 5. Consequently, it is not surprising that no shrinkage of the red region occurs as frequency is increased over the indicated range.

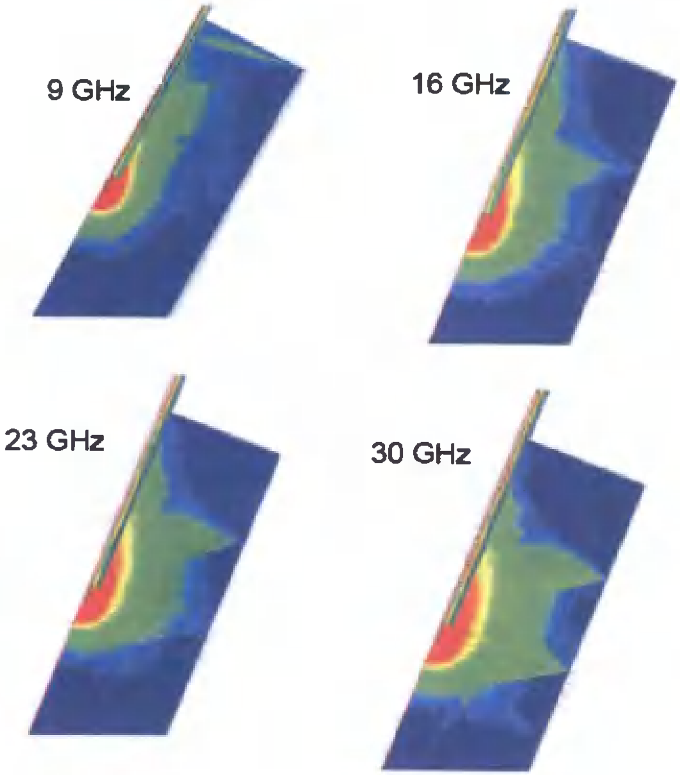


Fig. 5. Power loss density patterns in human breast fat at various frequencies (antenna length h = 0.9 mm).

## 3. EXPERIMENTS

A limited experimental verification of the HFSS simulation predictions was carried out at a frequency near 9 GHz. Chicken breast was used as a substitute for human muscle since at 9 GHz its electrical properties are similar ($\varepsilon_r \sim 40$, $\sigma \sim 10$ S/m) [11]. The experimental configuration is shown in Fig. 6. The mini co-axial cable that was inserted into the chicken breast was EZ Form copper-jacketed, semi-rigid mini cable with a characteristic impedance of 50 Ohms, average power rating of 3 Watts and attenuation of 1.9 dB per foot at 10 GHz [12]. The cable dimensions and dielectric permittivity were the same as were used in the HFSS simulations; the chosen length of the mini co-ax was 11 cm. To insert the mini-co-ax into the chicken breast, a channel was made using a hypodermic needle with a catheter. The needle and catheter were then removed and replaced by the mini co-ax. Alternate configurations which might be studied in the future include the following: 1) inserting the mini co-ax inserted into the catheter after the needle is removed; 2) building a mini co-ax into a hypodermic needle perhaps with the hollow needle functioning as the outer conductor of the co-axial transmission line. Option 2 would require a mechanism for inserting the antenna beyond the needle tip after the channel is formed by the needle.

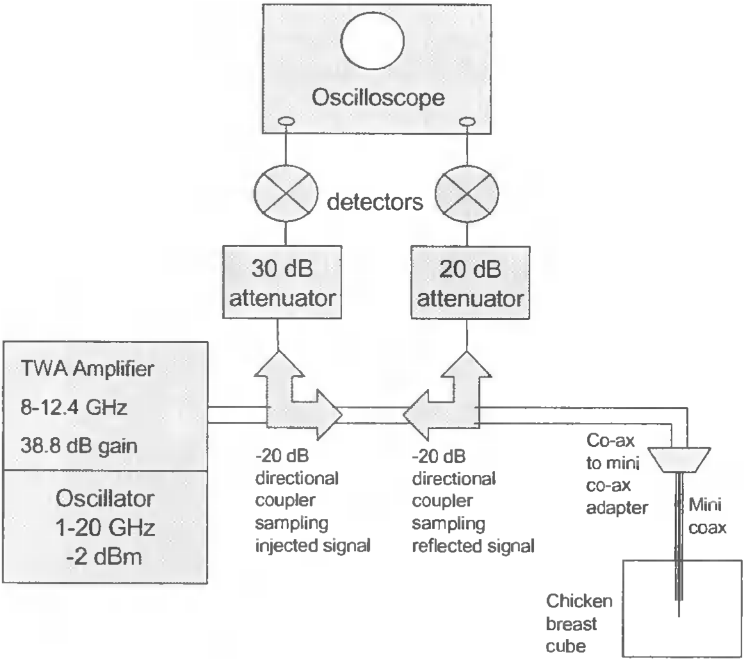


Fig. 6. Experimental configuration.

To minimize reflection of the microwaves from the chicken breast, the $S_{11}$ scattering parameter of the adapter and mini co-ax inserted into cold chicken breast was measured using a vector network analyzer which applied a signal power level of 100 milliwatts. The measurement was carried out for several lengths of the monopole antenna; viz., h = 0.9 mm, h = 2.5 mm and h = 5.0 mm. Results of these measurements are shown in Figure 7. Minimum reflections occur at specific values of frequency where impedance matching is optimum. The resonant frequencies are determined by such factors as the length of the mini-coaxial cable and shift only slightly with the length of the monopole antennas; however the depth of the absorption at each resonance is strongly affected by the antenna length. In the "hot" experiments, we chose the microwave frequency to obtain minimal reflection for each antenna length. For antenna length h = 0.9 mm, 9.27 GHz was chosen, and for h = 2.5 mm, 9.23 GHz was chosen. With h = 5.0 mm, the minimum value of $S_{11}$ is -16 dB; with h = 2.5 mm, the minimum value is – 15 dB; however, with h = 0.9 mm, the minimum value of $S_{11}$ is only – 7.5 dB, implying that with a source power of 10 Watts, only ~6 Watts would be deposited in the chicken breast. The difficulty of achieving a good impedance match when the antenna length becomes smaller than about 1 mm is clearly seen in Fig. 7. The choice h = 0.9 mm may be near the practical limit.

The chicken breast was irradiated with microwaves and then sliced open to determine the size of the region affected. Cell damage was indicated by the tissue changing color from pink to white. With 10 Watts of source power exposure times of only about 1 second were required. In Fig. 8, the interior of the chicken breast exposed by an axial cut can be seen after being heated with microwaves using an antenna of length h = 0.9 mm. In Fig. 9, the interior of the chicken breast exposed by an axial cut respectively can be seen after irradiation with microwaves using a longer antenna length h = 2.5 mm.

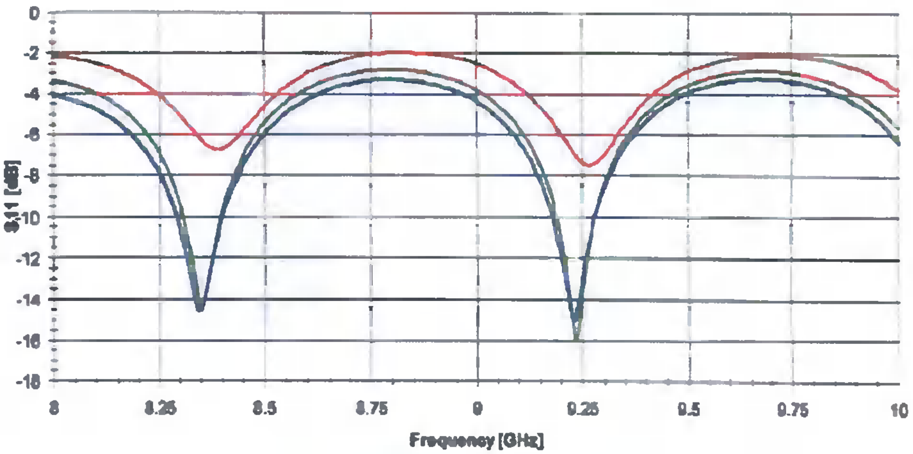


Fig. 7. $S_{11}$ measurements on adapter and mini co-ax [$S_{11}$ (dB) = 10 log ($P_{reflected}$ / $P_{in}$)] (red line is for h = 0.9 mm; blue line is for h = 2.5 mm; green line is for h = 5 mm).

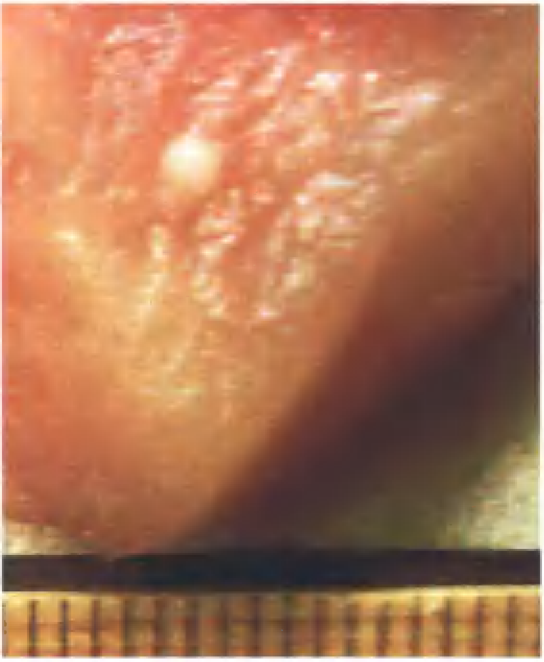


Fig. 8. Chicken breast interior exposed by axial cut after heating by microwaves using an antenna with h = 0.9 mm. Microwave pulse duration = 1.04 sec. f = 9.27 GHz. Extent of white region = 1.5 mm. Scale divisions are 1 mm in both Fig. 8 and Fig. 9.

It can be seen in Fig. 8, where h = 0.9 mm, that the affected white region has a diameter and an axial length which are both about 1.5 mm. Thus the affected region is approximately spherical in shape in reasonable agreement with the simulation results in Fig. 3b. In Fig. 9, where a longer antenna is used with h = 2.5 mm, the affected white region is larger having a length of about 5.5 mm and a maximum diameter of about 3.5 mm. This elongated shape of the heated region, is in reasonable agreement with the simulation results in Fig. 3a. Thus the experiment provides some validation of the HFSS simulation.

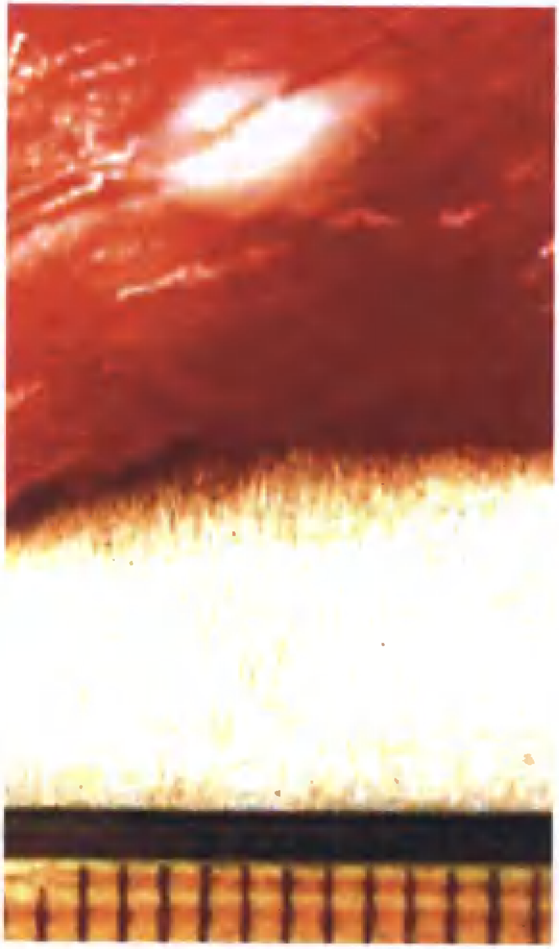


Fig. 9. Chicken breast interior exposed by axial cut after heating by microwaves using an antenna with h = 2.5 mm. Microwave pulse duration = 1.1 sec. f = 9.23 GHz. Extent of white region 5.5 mm long x 3.5 mm diameter.

## 4. CONCLUSIONS

Several significant conclusions can be drawn from the simulations and experiments described above. These conclusions are as follows:

- With a simple monopole antenna at the end of a miniaturized co-axial cable a very small, approximately spherical region of biological tissue can be preferentially heated with a microwave pulse. With a microwave frequency of ~9 GHz, and a monopole antenna length of 0.9 mm, the diameter of the spherical region is ~1.5 mm

- In tissue where the microwave absorption length is on the order of a millimeter in the operating frequency range, the diameter of the heated region can be decreased by raising the microwave frequency. For example, in human muscle tissue, as frequency is increased from 9 GHz to 30 GHz, the diameter of the heated sphere would shrink to approximately two-thirds of its original value.
- Hyperthermia of a small volume of biological tissue can be very rapid. With microwave pulse power of 10 Watts, hyperthermia of a spherical region with a diameter of 1.5 mm can be accomplished with pulse duration of 1 second or less. [Destruction of tumors occur at temperatures in the range 42 - 45 degrees Celsius; whereas, higher temperatures were reached in the present experiment when the chicken breast color changed from pink to white.]

Future studies should include more sophisticated modeling; for example, a region of malignant tissue could be defined inside a region of healthy tissue; the electrical properties of the two regions will differ. Also, the electrical properties of biological tissue are a function of temperature, and temperature will change as the microwave energy is absorbed; thus, the modeling should be dynamic and self-consistent, allowing for time dependence of the electrical properties of the tissue being heated. Such dynamic modeling has been incorporated into codes used to describe the sintering of ceramics [13, 14] and these codes should be adaptable to the case of heating biological material [15].

## ACKNOWLEDGEMENTS

The authors are indebted to Kerry Gaston and Harold Cox, M.D., who made them aware of the need for a minimally invasive therapy for tumors with diameters as small as one millimeter.

## BIBLIOGRAPHY

[1] Converse M., Bond E.J., Van Veen B.D., Hagness S.C., "A Computational Study of Ultra-Wideband Versus Narrowband Microwave Hyperthermia for Breast Cancer Treatment", *IEEE Trans. Microwave Theory and Technique* 54, pp. 2169-2180, 2006.

[2] Coughlin C.T., "Prospects for Interstitial Hyperthermia" in *Hyperthermia and Oncology*, vol. 3 (eds. M. Urano and E. Doyle, VSP, Utecht, The Netherlands), pp. 1-10, 1991.

[3] Robinson D.S., Parel J.-M., Denham D.B., Gonzalez-Cirre X., Manns F., Milne P.J., Schachner R.D., Herron A.J., Comander J., Hauptmann G., " Interstitial Laser Hyperthermia Model Development for Minimally Invasive Therapy of Breast Carcinoma", *Journal of the American College of Surgeons* 186, pp. 284-292, 1998.

[4] Trembly B.S., "The Effects of Driving Frequency and Antenna Length on Power-Deposition within a Microwave Antenna Array Used for Hyperthermia", *IEEE Trans. Biomed. Eng.* 32, pp. 152-157, 1985.

[5] Jerby E., Dikhtyar V., Aktushev O., Grosglick U., "The Microwave Drill", *Science* 298, 587-589, 2002.

[6] Dikhtyar V., Jerby E., "Fireball Ejection from a Molten Hot Spot to Air by Localized Microwaves", *Physical Review Letters* 96, article i.d. 045002, 5 pages, 2006.

[7] Bałzano Q., Foster K.R., Sheppard A.R., "Field and Temperature Gradients from Short Conductors in a Dissipative Medium", *Int. J. Antennas and Propagation*, article i.d. 57670, 8 pages, 2007.

[8] Barrett Alan H., Meyers Philip C., "Subcutaneous Temperatures: a Method of Noninvasive Sensing", *Science* 190, pp. 669-671, 1975.

[9] Data from the website of the Italian National Research Council, Institute for Applied Physics "Nello Carrara" – Florence (http://niremf.ifac.cnr.it/cgibin/tissprop/htmlclie/uniquery).

[10] HFSS is a product of Ansoft Corporation (http://www.ansoft.com/products/hf/hfss/).

[11] Miura N., Yagihara S., Mashimo S., "Microwave Dielectric Properties of Solid and Liquid Foods Investigated by Time-domain Reflectometry", *J. Food Engineering and Physical Properties* 68, pp.1396-1402, 2003.

[12] Complete specifications of the mini co-axial cable available on the website http://www.ezform.com/Products34.htm.

[13] Birnboim A., Gershon D., Calame J., Birman A., Carmel Y., Rodgers J., Levush B., Bykov Yu., Eremeev A.G., Dadon D., Martin L.P., Rozen M., Hutcheon R. "Comparative Study of Microwave Sintering of Zinc Oxide at 2.45, 30, and 83 GHz", J. Am. Ceram. Soc. 81, pp. 1493-501, 1998.

[14] Birnboim, Y. Carmel "Simulation of Microwave Sintering of Ceramic Bodies with CompleX Geometry", J. Am. Ceram. Soc. 82, pp. 3024-30, 1999.

[15] Birnboim A., private discussions.

# PERSPEKTYWY WEWNĄTRZTKANKOWEJ HIPERTERMII W USUWANIU ZMIAN RAKOWYCH Z UŻYCIEM PROMIENIOWANIA MIKROFALOWEGO W ZAKRESIE 9-30 GHz

## Streszczenie

Ten artykuł jest poświęcony metodzie wykorzystania energii promieniowania elektromagnetycznego z zakresu częstotliwości od 9 GHz do 30 GHz w uzyskaniu pożądanych skutków medycznych w małych objętościach materiału biologicznego. Zarówno symulacje numeryczne działania wypromieniowanej fali w tkance ludzkiej, jak i eksperymenty z tkanką kurczęcia (która ma podobne właściwości do tkanki ludzkiej), zostały wykonane i są tutaj omówione. Pokazano, że przy wykorzystaniu bardzo krotkiej jednopolowej anteny zainstalowanej na końcu miniaturowego współosiowego kabelka, wypromieniowanej mocy 10 W w impulsie o czasie trwania 1 s i częstotliwości fali 9 GHz, można wybraną objętość tkanki o kształcie kuli o średnicy 1,5 mm ogrzać do temperatury, w której zachodzi zniszczenie jej komórek. Ponadto z symulacji wynika, że jeżeli częstotliwość promieniowania mikrofalowego jest zwiększona do 30 GHz, to średnica podgrzanej kuli redukuje się do 1 mm. Zmiany rakowe w tkance ludzkiej są często diagnozowane wtedy, gdy mają one właśnie średnicę około 1 mm. Wyniki przedstawione w tym artykule mają charakter wstępny (rozpoznawczy) i zostały otrzymane w projekcie wykonanym przez studentów kończących magisterskie studia inżynierskie.

Słowa kluczowe: hipertermia mikrofalowa, wewnątrztkankowy aplikator, niewielkie zmiany rakowe, pasmo X, pasmo K

UNIWERSYTET TECHNOLOGICZNO-PRZYRODNICZY
IM. JANA I JĘDRZEJA ŚNIADECKICH W BYDGOSZCZY
ZESZYTY NAUKOWE NR 250
TELEKOMUNIKACJA I ELEKTRONIKA 11 (2008), 65-79

# OPIS LINIOWYCH UKŁADÓW I SYSTEMÓW CYFROWYCH ZA POMOCĄ SUMY SPLOTOWEJ A WŁASNOŚĆ ZANIKAJĄCEJ PAMIĘCI

Andrzej Borys, Zbigniew Zakrzewski

Wydział Telekomunikacji i Elektrotechniki
Uniwersytet Technologiczno-Przyrodniczy w Bydgoszczy
ul. Kaliskiego 7, 85-789 Bydgoszcz

*Streszczenie*: Ważną własnością, którą posiadają fizyczne układy i systemy analogowe (tj. czasu ciągłego) oraz cyfrowe (tj. czasu dyskretnego), jest tzw. własność zanikającej pamięci. Jej matematycznie precyzyjna definicja została po raz pierwszy podana przez Boyda i Chua. Udowodnili oni również, że opis liniowych układów i systemów cyfrowych za pomocą sumy splotowej jest możliwy wtedy i tylko wtedy, gdy posiadają one ww. własność. W tym artykule został omówiony i objaśniony dowód powyższego stwierdzenia, ze wszystkimi szczegółami nie opublikowanymi przez Boyda i Chua. Te szczegóły, omówienia i objaśnienia są o tyle istotne, że pozwalają na zrozumienie innego podejścia do problematyki określania warunków, które zapewniają istnienie opisu liniowych układów i systemów cyfrowych za pomocą sumy splotowej. Zaproponował je ostatnio Sandberg.

Słowa kluczowe: cyfrowe układy i systemy liniowe, pojęcie zanikającej pamięci, opis za pomocą sumy splotowej

## 1. WSTĘP

Bardzo wiele układów i systemów fizycznych można opisać za pomocą operatorów liniowych, to znaczy takich operatorów, które spełniają jednocześnie zasadę superpozycji (warunek addytywności) i warunek jednorodności. Oznacza to, że w tym przypadku spełnione jest następujące równanie:

$$G(\alpha x_1 + \beta x_2) = \alpha G x_1 + \beta G x_2, \qquad (1)$$

gdzie $G$ oznacza operator, $x_1$ i $x_2$ są dwoma różnymi sygnałami wejściowymi, a $\alpha$ i $\beta$ są dowolnymi liczbami rzeczywistymi.

O układach i systemach opisywanych za pomocą operatorów liniowych mówimy, że są to układy i systemy liniowe. W tej pracy ograniczamy się do rozpatrywania tylko cyfrowych układów i systemów liniowych, tj. pracujących z sygnałami czasu dyskretnego (sygnałami dyskretnymi).

Operatory liniowe (tak samo jak i nieliniowe), które służą do opisu fizycznych układów i systemów analogowych oraz cyfrowych (liniowych albo nieliniowych), posiadają różnego rodzaju własności, np. własność stacjonarności (niezależności od czasu) [1], własność tzw. zanikającej pamięci [1] itp.

W tym artykule koncentrujemy się na roli, jaką pełni własność zanikającej pamięci w opisie układu lub systemu cyfrowego (reprezentowanego przez operator liniowy) za pomocą sumy splotowej.

Idea pojęcia zanikającej pamięci nie jest nowa, występuje ona w wielu wcześniejszych pracach, m. in. w [2-5]. Jednakże po raz pierwszy matematycznie precyzyjną definicję tego pojęcia podali Boyd i Chua w pracy [1] i skutecznie wykorzystali w udowodnieniu kilku interesujących twierdzeń aproksymacyjnych. Między innymi udowodnili, że opis liniowych układów i systemów cyfrowych za pomocą sumy splotowej jest możliwy wtedy i tylko wtedy, gdy posiadają one ww. własność. W tej pracy zostanie omówiony i objaśniony ich dowód powyższego stwierdzenia, wraz z podaniem wszystkich szczegółów pominiętych w artykule [1]. Są one potrzebne w zrozumieniu innego podejścia do problematyki określania warunków, które zapewniają istnienie opisu liniowych układów i systemów cyfrowych za pomocą sumy splotowej. To nieco inne podejście pochodzi od Sandberga i zostało przez niego przedstawione m. in. w pracach [6], [7] i [8].

Na pierwszy rzut oka wydaje się, że każdy liniowy operator czasu dyskretnego (w skrócie, liniowy operator dyskretny) musi posiadać reprezentację w postaci sumy splotowej. Okazuje się jednak, że nie jest to prawda; pokazano to, między innymi, w wyżej wymienionych pracach [1] i [6].

Jednakże, jak pokazali Boyd i Chua [1], jeżeli dyskretny operator liniowy posiada własność tzw. zanikającej pamięci, to można go przedstawić w postaci sumy splotowej. Podobnie ma się rzecz z reprezentacją Sandberga [6] dla dyskretnego operatora liniowego. Jeżeli drugi składnik w tej reprezentacji, będący granicą pewnego wyrażenia dla dyskretnego czasu zdążającego do nieskończoności, równa się zero, to operator posiada opis w postaci sumy splotowej.

Zachodzące związki pomiędzy ww. podejściami do problemu reprezentacji dyskretnego operatora liniowego za pomocą sumy splotowej zostaną przedstawione w następnym artykule.

## 2. DEFINICJE WŁASNOŚCI STACJONARNOŚCI I ZANIKAJĄCEJ PAMIĘCI DLA OPERATORÓW DYSKRETNYCH

W celu sformułowania zasadniczego twierdzenia, omawianego w następnym rozdziale, oraz przeprowadzenia jego dyskusji, konieczne jest wprowadzenie kilku oznaczeń i definicji. Niech zatem symbol $\mathbb{Z}$ oznacza zbiór liczb całkowitych, natomiast $\mathbb{Z}_+$ i $\mathbb{Z}_-$ zbiory, odpowiednio, liczb całkowitych nieujemnych i liczb całkowitych niedodatnich. Ponadto, niech $l^\infty(\cdot)$ oznacza przestrzeń sygnałów dyskretnych będących sekwencjami (ciągami) ograniczonymi z normą

$$\|x\| \overset{df}{=} \sup|x(k)|, \tag{2}$$

gdzie $k$ oznacza czas dyskretny. W dalszej części artykułu korzystamy z następującej notacji: $l^\infty(\mathbb{Z}_-)$, gdy $k \in \mathbb{Z}_-$, $l^\infty(\mathbb{Z}_+)$, gdy $k \in \mathbb{Z}_+$ oraz $l^\infty(\mathbb{Z}) = l^\infty$, gdy $k \in \mathbb{Z}$.

Operatorem opóźniającym o $\tau$ jednostek czasu, $\tau \in \mathbb{Z}$, nazwiemy taki operator, którego równanie definicyjne ma następującą postać:

$$(U_{\tau}x)(k) \overset{df}{=} x(k-\tau)\,. \tag{3}$$

Natomiast operator $G$ nazwiemy operatorem stacjonarnym, tj. niezależnym od czasu (ang. *time-invariant*), gdy spełnia równanie

$$U_{\tau}G = GU_{\tau}\,, \tag{4}$$

dla wszystkich $\tau \in \mathbb{Z}$.

Ponadto, operator $G$ będziemy nazywać operatorem skutkowo-przyczynowym (ang. *causal*), gdy równość sygnałów wejściowych $x(\tau) = v(\tau)$ dla $\tau \leq k$, $\tau, k \in \mathbb{Z}$, będzie pociągać za sobą następującą równość:

$$(Gx)(k) = (Gv)(k)\,. \tag{5}$$

Powyższe definicje dyskretnych operatorów: stacjonarnego (niezależnego od czasu) oraz skutkowo-przyczynowego – a także podana poniżej definicja operatora posiadającego własność zanikającej pamięci (ang. *fading memory*) – są słuszne niezależnie od tego czy rozpatrywany operator jest liniowy, czy nieliniowy. Więcej szczegółów na temat tych definicji oraz zastosowania w aproksymacji nieliniowych systemów cyfrowych można znaleźć w pracach [1] i [9].

Definicja zanikającej pamięci, tak jak została ona sformułowana przez Boyda i Chua w pracy [1], brzmi następująco:

Operator niezależny od czasu $G\colon l^{\infty} \to l^{\infty}$ posiada zanikającą pamięć na podprzestrzeni $K \subset l^{\infty}$, jeżeli istnieje taka sekwencja $w\colon \mathbb{Z}_{+} \to (0,1>$, $\lim_{k\to\infty} w(k) = 0$, że dla każdego $x \in K$ i każdego $\varepsilon > 0$ istnieje takie $\delta > 0$, że dla wszystkich $v \in K$ zachodzi

$$\sup_{k\leq 0}\left|x(k)-v(k)\right|w(-k)<\delta \quad \to \quad \left|(Gx)(0)-(Gv)(0)\right| < \varepsilon\,. \tag{6}$$

Nierówność (6) oznacza, że wartość operatora $G$ dla bieżącej chwili czasowej (tutaj $k=0$) w coraz mniejszym stopniu zależy od wartości sygnału z przeszłości (to jest gdy $k \to -\infty$).

## 3. TWIERDZENIE BOYDA-CHUA O REPREZENTACJI ZA POMOCĄ SUMY SPLOTOWEJ

W pracy [1] zostało podane następujące twierdzenie o reprezentacji dyskretnego operatora liniowego za pomocą sumy splotowej.

Twierdzenie (Boyd i Chua [1]). Operator $G\colon l^{\infty} \to l^{\infty}$ jest operatorem liniowym, niezależnym od czasu oraz posiadającym zanikającą pamięć wtedy i tylko wtedy, gdy może być przedstawiony w postaci sumy splotowej

$$y(k) = Gx(k) = h(k) \otimes x(k) = \sum_{n=0}^{\infty} h(n)x(k-n)\,, \tag{7}$$

gdzie $h \in l^1(\mathbb{Z}_+)$, ($l^1(\mathbb{Z}_+)$ oznacza przestrzeń ciągów liczbowych z normą $\|h\|_1 = \sum_{k=0}^{\infty} |h(k)| < \infty$. Symbol $\otimes$ w równaniu (7) oznacza sumę splotową.

Dowód. Załóżmy, że suma splotowa dana wzorem (7), gdzie $h \in l^1(\mathbb{Z}_+)$, istnieje. Oznaczmy ją symbolem $G$. Następnie wykażemy, że takie $G$: $l^\infty \to l^\infty$ jest operatorem liniowym, niezależnym od czasu i posiadającym własność zanikającej pamięci.

W celu wykazania liniowości przekształcenia $G$ założymy sygnał wejściowy w postaci sumy dwóch sygnałów $x_1(k)$ i $x_2(k)$ pomnożonych przez liczby rzeczywiste, odpowiednio, $\alpha$ i $\beta$. Podstawiając taki sygnał do wzoru (7), możemy napisać

$$\begin{aligned} G(\alpha x_1 + \beta x_2)(k) &= \sum_{n=0}^{\infty} h(n)\{\alpha x_1(k-n) + \beta x_2(k-n)\} = \\ &= \alpha \sum_{n=0}^{\infty} h(n) x_1(k-n) + \beta \sum_{n=0}^{\infty} h(n) x_2(k-n) = \alpha G x_1(k) + \beta G x_2(k). \end{aligned} \tag{8}$$

Ze wzoru (8) wynika, że operator $G$ spełnia warunki addytywności i jednorodności. A zatem, zgodnie z równaniem definicyjnym operatora liniowego (1), przekształcenie $G$ należy do tej klasy operatorów.

Pokażemy teraz, że operator dany wzorem (7) jest operatorem niezależnym od czasu. W tym celu zauważmy najpierw, że

$$h \otimes (U_\tau x) = h(k) \otimes x(k-\tau) = \sum_{n=0}^{\infty} h(n) x(k-\tau-n). \tag{9}$$

gdzie operator $U_\tau$ jest operatorem opóźniającym sygnał o $\tau$ jednostek czasowych, natomiast operator $G$, o którym mowa w równaniu (4), jest reprezentowany tutaj przez sekwencję $h(k)$ zawierającą próbki tzw. odpowiedzi impulsowej układu (systemu) opisywanego za pomocą sumy splotowej.

Z drugiej zaś strony mamy

$$U_\tau(h \otimes x) = \sum_{n=0}^{\infty} h(n) x(k-\tau-n). \tag{10}$$

Porównując ze sobą wzory (9) i (10), dochodzimy do wniosku, że zachodzi

$$U_\tau(h \otimes x) = h \otimes (U_\tau x) \tag{11}$$

a zatem operator $G$ jest operatorem niezależnym od czasu, zgodnie z równaniem definicyjnym (4).

Weźmy teraz pod uwagę sekwencję wagową o wyrazie ogólnym $w(k)$ danym następującym wzorem:

$$w(k) = \|h\|_1^{-1/2} \left\{ \sum_{n=k}^{\infty} |h(n)| \right\}^{1/2}. \tag{12}$$

(Sekwencja $w(k)$ dana wzorem (12) jest sekwencją malejącą, przyjmującą wartości z przedziału $(0,1>$ i $\lim_{k\to\infty} w(k)=0$. Przy okazji zauważmy, że możliwe są również inne wybory dla $w(k)$, spełniające powyższe własności).

Pokażemy następnie, że udowodnienie, że $G$ posiada zanikającą pamięć sprowadza się do udowodnienia, że zachodzi następująca nierówność:

$$S=\sum_{n=0}^{\infty}|h(n)|w^{-1}(n)<\infty\,. \tag{13}$$

W tym celu rozpatrzymy wyrażenie

$$|Gx(0)-Gv(0)| = \left|\sum_{n=0}^{\infty} h(n)(x(0-n)-v(0-n))\right| \le \sum_{n=0}^{\infty}|h(n)||x(-n)-v(-n)|\,. \tag{14}$$

Zauważmy dalej, że w definicji zanikającej pamięci rozpatrujemy takie sygnały, dla których zachodzi

$$\sup_{n\ge 0}|x(-n)-v(-n)|\,w(n)<\delta\,, \tag{15}$$

skąd

$$|x(-n)-v(-n)|<\frac{\delta}{w(n)},\ n\ge 0\,. \tag{16}$$

Podstawiając (16) w (14), otrzymuje się:

$$|Gx(0)-Gv(0)|<\delta\sum_{n=0}^{\infty}|h(n)|w^{-1}(n)=\varepsilon\,. \tag{17}$$

Z nierówności (17) wynika, że przy spełnionej nierówności (13), można zawsze dobrać $\delta$ tak, aby uczynić $|Gx(0)-Gv(0)|$ dowolnie małym. Podsumowując zatem, można powiedzieć, że prawdziwość nierówności (17) jest równoważna faktowi posiadania przez operator $G$ zanikającej pamięci. Do udowodnienia pozostaje tylko wykazanie, że rzeczywiście (13) zachodzi.

W tym celu zdefiniujmy zmienną $m(n)$ w następujący sposób:

$$m(n)=\sum_{k=n}^{\infty}|h(k)|\,. \tag{18}$$

Zauważmy, że z założenia $h\in l^1(\mathbb{Z}_+)$ wynika $m(0)<\infty$. W dalszych rozważaniach będziemy też korzystać z faktu, że $m(n)>0,\ n=0,1,2,\ldots$.

Podzielmy teraz sumę $S$ ze wzoru (13) przez $m(0)$ i wprowadźmy zmienną $m(n)$. W ten sposób otrzymamy:

$$S'=\frac{1}{m(0)}\sum_{n=0}^{\infty}\frac{|h(n)|\left\{\sum\limits_{k=0}^{\infty}|h(k)|\right\}^{1/2}}{\left\{\sum\limits_{k=n}^{\infty}|h(k)|\right\}^{1/2}}=\{m(0)\}^{-1/2}\sum_{n=0}^{\infty}\frac{|h(n)|}{\{m(n)\}^{1/2}}\,. \tag{19}$$

Zauważmy dalej, że $|h(n)|$ można napisać w postaci:

$$|h(n)| = \sum_{k=n}^{\infty} |h(k)| - \sum_{k=n+1}^{\infty} |h(k)| \,. \tag{20}$$

A zatem, korzystając ze wzorów (18) i (20) we wzorze (19), mamy:

$$S' = \{m(0)\}^{-1/2} \sum_{n=0}^{\infty} \frac{m(n) - m(n+1)}{\{m(n)\}^{1/2}} \,. \tag{21}$$

Wzór (21) możemy też rozpisać w następujący sposób:

$$\begin{aligned} S' &= \{m(0)\}^{-1/2} \left\{ \frac{m(0)}{m(0)^{1/2}} + \frac{-m(1)}{m(0)^{1/2}} + \frac{m(1)}{m(1)^{1/2}} + \dots \right\} = \\ &= 1 + \{m(0)\}^{-1/2} \sum_{n=0}^{\infty} m(n+1)\left(m(n+1)^{-1/2} - m(n)^{-1/2}\right). \end{aligned} \tag{22}$$

Zauważmy dalej, że

$$\begin{aligned} & m(n+1)\left(m(n+1)^{-1/2} - m(n)^{-1/2}\right) = m(n+1)^{1/2} - m(n+1)^{1/2}\frac{m(n+1)^{1/2}}{m(n)^{1/2}} = \\ & = m(n)^{1/2} - m(n)^{1/2} + m(n+1)^{1/2} - m(n+1)^{1/2}\frac{m(n+1)^{1/2}}{m(n)^{1/2}} = \\ & = m(n)^{1/2} - m(n+1)^{1/2}\left[\frac{m(n+1)^{1/2}}{m(n)^{1/2}} + \frac{m(n)^{1/2}}{m(n+1)^{1/2}} - 1\right]. \end{aligned} \tag{23}$$

Pokażemy teraz, że dla wyrażenia w nawiasach kwadratowych w równaniu (23) zachodzi

$$\frac{m(n+1)^{1/2}}{m(n)^{1/2}} + \frac{m(n)^{1/2}}{m(n+1)^{1/2}} - 1 \ \geq\ 1 \,. \tag{24}$$

W tym celu przeniesiemy jedynkę występującą po lewej stronie nierówności (24) na jej prawą stronę. Następnie podniesiemy do kwadratu wyrażenia po obydwu stronach nierówności. W wyniku otrzymamy:

$$\frac{m(n+1)}{m(n)} + \frac{m(n)}{m(n+1)} \ \geq\ 2 \,. \tag{25}$$

Zauważmy dalej, że zachodzi:

$$0 < m(n+1) \ \leq m(n) \,. \tag{26}$$

A zatem możemy napisać:

$$m(n) = m(n+1) + \Delta \,, \tag{27}$$

gdzie:

$$m(n) - m(n+1) \ \geq\ 0 \,. \tag{28}$$

Podstawiając (27) do nierówności (25), otrzymamy:

$$1+\frac{\Delta}{m(n+1)}+\frac{1}{1+\frac{\Delta}{m(n+1)}} \geq 2. \tag{29}$$

Przekształcając zaś następnie nierówność (29), mamy:

$$1+\frac{\Delta}{m(n+1)} \geq 2-\frac{1}{1+\frac{\Delta}{m(n+1)}} = \frac{1+\frac{2\,\Delta}{m(n+1)}}{1+\frac{\Delta}{m(n+1)}} \tag{30}$$

i z (30) otrzymujemy:

$$\left(1+\frac{\Delta}{m(n+1)}\right)^2 = 1+\frac{2\,\Delta}{m(n+1)}+\left(\frac{\Delta}{m(n+1)}\right)^2 \geq 1+\frac{2\,\Delta}{m(n+1)} \tag{31}$$

Ostatecznie z nierówności (31) dostajemy:

$$\left(\frac{\Delta}{m(n+1)}\right)^2 \geq 0. \tag{32}$$

to znaczy nierówność, która jest prawdziwa. Zatem prawdziwa jest również nierówność, od której wyszliśmy, to jest (24).

Zastosowanie nierówności (24) w (23) pozwala na napisanie

$$m(n+1)\left(m(n+1)^{-1/2}-m(n)^{-1/2}\right) \leq m(n)^{1/2}-m(n+1)^{1/2}. \tag{33}$$

Uwzględniając następnie (33) w wyrażeniu (22) na sumę $S'$, otrzymamy:

$$\begin{aligned} S' &\leq 1+\{m(0)\}^{-1/2}\sum_{n=0}^{\infty}\left(m(n)^{1/2}-m(n+1)^{1/2}\right)= \\ &= 1+\{m(0)\}^{1/2}\left[m(0)^{1/2}-m(1)^{1/2}+m(1)^{1/2}-m(2)^{1/2}+\ldots\right]=1+1=2 \end{aligned} \tag{34}$$

ponieważ

$$\lim_{n\to\infty} m(n)^{1/2}=0. \tag{35}$$

Na koniec biorąc pod uwagę (34) i fakt, że $m(0)<\infty$, możemy napisać

$$S \leq 2\,m(0) < \infty. \tag{36}$$

W ten sposób, dowodząc prawdziwości nierówności (13), dowiedliśmy, że operator $G$ posiada także własność zanikającej pamięci.

Teraz przeprowadzimy dowód twierdzenia w stronę odwrotną. To znaczy założymy, że rozpatrywany operator dyskretny $G$ jest operatorem liniowym, niezależnym od czasu i posiadającym własność zanikającej pamięci. Udowadniać zaś będziemy, że ten operator posiada reprezentację w postaci sumy splotowej, tj. że zachodzi $Gx(k) = h(k) \otimes x(k)$ dla wszystkich $x \in l^{\infty}$.

Niech $h$ będzie odpowiedzią układu (systemu) opisywanego powyższym operatorem $G$ na jednostkową „próbkę", tj. $h(n) = Ge(n)$, gdzie $e(n) = \delta_{n,0}$, z $\delta_{n,0}$ oznaczającym symbol Kroneckera (z indeksami ze zbioru liczb całkowitych).

Pokażemy, że $h \in l^{1}(\mathbb{Z}_{+})$ (zauważmy, że w tym momencie wiemy tylko tyle, że $h \in l^{\infty}(\mathbb{Z}_{+})$, gdyż operator $G$ jest operatorem $G$: $l^{\infty} \to l^{\infty}$).

Dalej, niech $F$ będzie funkcjonałem stowarzyszonym z operatorem $G$ poprzez równanie

$$Fx \overset{\mathrm{df}}{=} Gx_{c}(0)\,, \tag{37a}$$

gdzie $x \in l^{\infty}(\mathbb{Z}_{-})$, a sekwencja $x_{c}(k)$ dana jest wzorem

$$x_{c}(k) \overset{\mathrm{df}}{=} \begin{cases} x(k) & \text{dla} \quad k \le 0 \\ 0 & \text{dla} \quad k > 0 \end{cases} \tag{37b}$$

Pokażemy teraz, że dla wszystkich $x \in l^{\infty}(\mathbb{Z}_{-})$ istnieje takie $M < \infty$, że można napisać

$$|Fx| \;\le\; M\|x\|_{w}\,, \tag{38}$$

gdzie $\|x\|_{w}$ oznacza ważoną normę supremum sygnału dyskretnego, określoną wzorem

$$\|x\|_{w} \overset{df}{=} \sup_{k \le 0} |x(k)w(-k)|\,. \tag{39}$$

z sekwencją wagową $w(k)$ o właściwościach podanych w definicji zanikającej pamięci.

W celu wykazania prawdziwości nierówności (38) skorzystamy z przyjętego założenia, że operator $G$ jest operatorem liniowym posiadającym własność zanikającej pamięci. I rozważymy, w pierwszej kolejności, zbiór wszystkich sekwencji $x$, dla których zachodzi

$$0 < \delta_{1} < \|x\|_{w} < \delta_{2}\,, \tag{40}$$

gdzie $\delta_1$ i $\delta_2$ są liczbami rzeczywistymi dodatnimi. Zbiór ten oznaczymy przez $X_{M}$.

Następnie przyjmiemy w definicji zanikającej pamięci (6), że sekwencja $v(k) \equiv 0$ dla $k \in \mathbb{Z}$ (po prostu wybieramy dla naszych dalszych rozważań, w definicji (6), sekwencję $v(k)$ o wszystkich elementach będących zerami). Zauważmy, że wtedy możemy napisać

$$\sup_{k \le 0} |x(k)|\, w(-k) < \delta \quad \to \quad |(Gx_{c})(0)| = |Fx| < \varepsilon\,. \tag{41}$$

W (41) wykorzystano również fakt, że z liniowości operatora $G$ wynika $(Gv)(0) = 0$, gdyż $G(\alpha v) = G(v) = \alpha G(v)$, skąd przy założeniu $\alpha \neq 0 \;\to\; G(v) = 0$, a zatem również $(Gv_{c})(0) = 0$.

Implikacja (41) dla pewnego odpowiednio dobranego $\varepsilon$ zachodzi dla wszystkich sekwencji ze zbioru $X_M$. to znaczy mamy:

$$\|x\|_w < \delta_2 \quad \rightarrow \quad |Fx| < \varepsilon \tag{42}$$

Zatem możemy stałą $\varepsilon$ zapisać w następującej postaci:

$$\varepsilon = M_i \|x_i\|_w, \quad x_i \in X_M \tag{43a}$$

skąd:

$$M_i = \frac{\varepsilon}{\|x_i\|_w} < \frac{\varepsilon}{\delta_1} = M < \infty \tag{43b}$$

co pozwala napisać

$$|Fx| < M\|x\|_w. \tag{44}$$

dla każdego $x \in X_M$. Oznacza to, że w przypadku zbioru $X_M$ nierówność (38) jest spełniona.

Zauważmy dalej, że każdą sekwencję $x_a \notin X_M$ (oprócz sekwencji zerowej, tj. takiej, której wszystkie wyrazy są równe tożsamościowo zero) można przedstawić w następujący sposób:

$$x_a = \alpha\, x_M\,, \quad \alpha \neq 0 \tag{45}$$

gdzie $\alpha$ oznacza liczbę rzeczywistą, a $x_M$ jest pewną sekwencją ze zbioru $X_M$. Przypadek z sekwencją zerową (z przestrzeni $l^{\cdot}(\mathbb{Z}\ )$ ) będziemy rozpatrywać oddzielnie.

Korzystając z nierówności (44) i liniowości operatora $G$, możemy napisać

$$|\alpha||Fx_M| = |F(\alpha x_M)| < |\alpha| M\|x_M\|_w = M\|\alpha x_M\|_w \tag{46}$$

Następnie wykorzystując w nierówności (46) fakt, że sekwencja nie należąca do zbioru $X_M$ może być zapisana w postaci (45), mamy

$$|Fx_a| < M\|x_a\|_w\,. \tag{47}$$

Z powyższego wynika, że nierówność (38) jest spełniona również dla sekwencji nie objętych nierównością (40) – z wyjątkiem sekwencji zerowej, którą zajmiemy się teraz. Przyporządkujmy tej sekwencji oznaczenie $x_0$ i zauważmy, że wykazaliśmy już (przy wyprowadzeniu nierówności (41)), że zachodzi $G(x_{0e}) = 0$ dla sekwencji zerowej $x_{0e}$ z przestrzeni $l$ . Z tego i zależności (37a) wynika, że $Fx_0 = 0$, a więc możemy napisać

$$|Fx_0| = |0| = 0 \leq M\|x_0\|_w = M \cdot 0 = 0\ . \tag{48}$$

A zatem nierówność (38) jest spełniona również dla sekwencji zerowej z przestrzeni $l^{\cdot}(\mathbb{Z}\ )$.

Zdefiniujemy teraz dla dowolnego sygnału dyskretnego $x:\mathbb{Z}\rightarrow\mathbb{R}$ sekwencję $x_N(k)$ (z przestrzeni $l^\infty(\mathbb{Z}_-)$ ) w następujący sposób:

$$x_N(k)=\begin{cases}x(k) & -N\leq k\leq 0\\ 0 & k<-N\end{cases} \tag{49a}$$

gdzie $N$ oznacza liczbę całkowitą nieujemną, a $x(k)\in l^\infty(\mathbb{Z})$. Ponadto, określimy stowarzyszoną z $x_N(k)$ sekwencję $x_{Ne}(k)$ z przestrzeni $l^\infty(\mathbb{Z})$ jako:

$$x_{Ne}(k)\overset{\text{df}}{=}\begin{cases}x_N(k) & \text{dla} \quad k\leq 0\\ 0 & \text{dla} \quad k>0\end{cases} \tag{49b}$$

Następnie skorzystamy z własności liniowości i niezależności od czasu operatora $G$: $l^\infty\rightarrow l^\infty$, aby pokazać, że zachodzi równość

$$Fx_N=\sum_{n=0}^{N}h(n)x_N(-n)=\sum_{n=0}^{N}h(n)x(-n)\,. \tag{50}$$

Zauważmy, że w zależności (50) rozpatrujemy chwilę czasową $k=0$, gdyż dla tej właśnie chwili czasowej mamy zgodnie z definicją obliczyć wartość funkcjonału, tj. $Fx_N=Gx(0)$. Ponadto, zauważmy, że zdefiniowaną przed chwilą sekwencję $x_N(k)$ można przedstawić w następujący sposób:

$$x_N(k)=\sum_{n=0}^{-N}x(n)\delta_{k,n}\,, \tag{51}$$

gdzie indeks $n$ w symbolu Kroneckera oznacza moment przyłożenia impulsu jednostkowego (jednostkowej „próbki").

Zauważmy również, że odpowiedź systemu na pobudzenie w postaci impulsu jednostkowego przyłożonego w chwili czasowej $k=(n=0)$ można napisać w postaci

$$(G\delta_{k,0})(k)=\sum_{l=0}^{\infty}h(l)\delta_{k,l}\,. \tag{52}$$

Jeżeli impuls jednostkowy jest przyłożony w chwili czasowej $k=(n\neq 0)$, to zgodnie z definicją niezależności od czasu operatora $G$, odpowiedź układu (systemu) opisywanego przez ten operator można przedstawić tak

$$(G\delta_{k,n})(k)=(G\delta_{k-n,0})(k-n)\,, \tag{53}$$

gdzie impuls jednostkowy po prawej stronie równości (53) oznacza impuls przyłożony w zerowej chwili czasowej. Należy też zwrócić uwagę, że odpowiedź układu (systemu) po prawej stronie równości (53) jest opóźniona o $n$ jednostek czasowych.

Wykorzystując teraz zależności (49a), (49b), (51), (52) i (53) oraz fakt, że operator $G$ jest operatorem liniowym, możemy napisać:

$$\begin{aligned}(Gx_{N_e})(k) &= G\left(\sum_{n=0}^{-N} x(n)\delta_{k,n}\right) = \sum_{n=0}^{-N} x(n)(G\delta_{k,n})(k) = \\ &= \sum_{n=0}^{-N} x(n)(G\delta_{k-n,0})(k-n) = \sum_{n=0}^{-N} x(n)\sum_{l=0}^{\infty} h(l)\delta_{k-n,l}\end{aligned} \tag{54}$$

Zauważmy, że z ostatniego równania dla chwili czasowej $k = 0$ wynika:

$$(Gx_{N_e})(0) = \sum_{n=0}^{-N} x(n)\sum_{l=0}^{\infty} h(l)\delta_{-n,l} \tag{55}$$

W powyższym równaniu druga suma wnosi coś do wartości wyrażenia tylko wtedy, gdy $-n = l$. Uwzględniając to, możemy (55) przepisać w postaci

$$(Gx_{N_e})(0) = \sum_{n=0}^{-N} x(n)h(-n) \tag{56}$$

Podstawiając następnie $n = -l$ w równaniu (56) oraz biorąc pod uwagę równania (37a) i (37b), otrzymuje się ostatecznie

$$Fx_N = (Gx_{N_e})(0) = \sum_{l=0}^{N} h(l)x(-l) \tag{57a}$$

lub

$$Fx_N = (Gx_{N_e})(0) = \sum_{l=0}^{N} h(l)x_N(-l) \tag{57b}$$

Weźmy teraz pod uwagę sekwencję

$$x_N(-k) = \begin{cases} w(k)^{-1}\operatorname{sign}\left(h(k)\right) & \text{dla} \quad 0 \le k \le N \\ 0 & \text{dla} \quad k > N \end{cases} \tag{58}$$

z przestrzeni $x \in l^{\infty}(\mathbb{Z}_-)$ i podstawmy ją do wzoru (57b). W wyniku otrzymamy:

$$Fx_N = \sum_{l=0}^{N} \left|h(l)\right| w(l)^{-1} \tag{59}$$

Uwzględniając następnie nierówność (38) w (59), możemy napisać:

$$\begin{aligned}\sum_{l=0}^{N} \left|h(l)\right| w(l)^{-1} \;\le\; M \sup_{-N\le k'\le 0} \left|w(-k')^{-1}\operatorname{sign}\left(h(-k')\right) w(-k')\right| = \\ = M \sup_{-N\le k'\le 0} \left|\operatorname{sign}\left(h(-k')\right)\right| = M\end{aligned} \tag{60}$$

skąd mamy:

$$\sum_{l=0}^{N} \left|h(l)\right| w(l)^{-1} \;\le\; M \tag{61}$$

dla wszystkich $N$. Oznacza to, że $hw^{-1} \in l^1(\mathbb{Z}_+)$, co pociąga za sobą $h \in l^1(\mathbb{Z}_+)$. Implikacja ta wynika z twierdzenia o porównywaniu szeregów [10], gdyż w naszym przypadku mamy

$$\frac{|h(k)|}{w(k)} \geq |h(k)| \quad , \quad \text{przy} \quad w(k) \in (0,\ 1>. \tag{62}$$

Na marginesie zauważmy, że fakt, że $h \in l^1(\mathbb{Z}_-)$ można wykazać w bardziej elegancki sposób, jeżeli się przyjmie w wyprowadzeniu powyżej sekwencję

$$x(-k) = \begin{cases} \operatorname{sign}\big(h(k)\big) & \text{dla} \quad k \geq 0 \\ 0 & \text{dla} \quad k < 0 \end{cases} \tag{63}$$

(z przestrzeni $l^{\infty}(\mathbb{Z})$), zamiast sekwencji danej wzorem (58). Dla tej sekwencji wziętej pod uwagę we wzorach (49a) i (49b), otrzymamy ze wzoru (57a)

$$Fx_N = (Gx)(0) = \sum_{l=0}^{N} |h(l)|. \tag{64}$$

Następnie wykorzystując definicję (37a) i nierówność (38) w (64), mamy

$$\sum_{l=0}^{N} |h(l)| \leq M \sup_{-N \leq k' \leq 0} \left|\operatorname{sign}\big(h(-k'\big) w(-k')\right| \leq M \tag{65}$$

dla każdego $N$, skąd bezpośrednio wynika, że $h \in l^1(\mathbb{Z}_-)$.

Pokażemy teraz, że istnieje reprezentacja operatora $G$ w postaci sumy splotowej. W tym celu zauważymy, że dla każdego $x \in l^{\infty}(\mathbb{Z}\_)$ z nierówności (38) wynika, że zachodzi następująca nierówność:

$$|Fx - Fx_N| \leq M \|x - x_N\|_w \leq M \sup_{k<-N} \big(|x(k)|\, w(-k)\big). \tag{66}$$

A ponieważ sekwencja $x(k)$ jest sekwencją ograniczoną, to znaczy

$$\sup_{k<-N} |x(k)| \leq \beta, \tag{67a}$$

gdzie $\beta$ oznacza pewną stałą (o skończonej wartości), więc możemy napisać

$$\sup_{k<-N} \big(|x(k)|\, w(-k)\big) \leq \beta \cdot w(N+1). \tag{67b}$$

Wykorzystując następnie (67b) w nierówności (66), otrzymamy

$$|Fx - Fx_N| \leq M \cdot \beta \cdot w(N+1) \to 0, \text{ gdy } N \to 0. \tag{68}$$

I ostatecznie, z zależności (57a) i nierówności (68), mamy

$$Fx = \lim_{N \to \infty} Fx_N = \sum_{n=0}^{\infty} h(n) x(-n). \tag{69}$$

Zauważmy też, że $h(\cdot)x(-\cdot) \in l^1(\mathbb{Z}_+)$, co można wykazać podstawiając do nierówności (38) sekwencję w postaci

$$x'(-k) = |x(-k)| \operatorname{sign}(h(k)) \quad \text{dla} \quad k \geq 0, \tag{70}$$

gdzie $x$ i $x' \in l^\infty(\mathbb{Z}_-)$. Wykorzystując ponadto (69), w wyniku otrzymamy

$$Fx' = \sum_{n=0}^{\infty} h(n)x'(-n) = \sum_{n=0}^{\infty} |h(n)||x(-n)| \leq M \|x(-n)\|_u . \tag{71}$$

Z faktu, że $\|x(-n)\| < \infty$ wynika w sposób natychmiastowy, że zachodzi również $\|x(-n)\|_u < \infty$. Przy wykorzystaniu (71) pozwala to napisać

$$\sum_{n=0}^{\infty} |h(n)||x(-n)| < \infty \tag{72}$$

a zatem rzeczywiście mamy $h(\cdot)x(-\cdot) \in l^1(\mathbb{Z}_+)$.

Zdefiniujmy teraz operator obcinający $P: l^\infty(\mathbb{Z}) \to l^\infty(\mathbb{Z}_-)$ w następujący sposób:

$$Px(k) = x(k) \quad \text{dla} \quad k \leq 0 \tag{73}$$

Jak widać z zależności (73), operator ten odwzorowuje element $x$ przestrzeni $l^\infty(\mathbb{Z})$ w element $Px$ przestrzeni $l^\infty(\mathbb{Z}_-)$.

W [1] i [9] pokazano, że każdy nieliniowy (a zatem, w szczególności, również liniowy $G$) dyskretny operator niezależny od czasu oraz posiadający zanikającą pamięć może być odzyskany ze stowarzyszonego z nim funkcjonału $F$ danego wzorem (37a). Zależność, która wiąże ze sobą $G$ i $F$ ma postać

$$Gx(k) = FPU_{-k}x . \tag{74}$$

Korzystając z (3), (73) i (69) w (74), można napisać

$$Gx(k) = \sum_{n=0}^{\infty} h(n)PU_{-k}x(-n) = \sum_{n=0}^{\infty} h(n)Px(k-n) = \sum_{n=0}^{\infty} h(n)x(k-n) \tag{75}$$

Zależność (75) pokazuje, że rzeczywiście dyskretny operator liniowy $G$, niezależny od czasu oraz posiadający własność zanikającej pamięci posiada reprezentację w postaci sumy splotowej (7), co kończy dowód twierdzenia.

## 4. WNIOSKI I PODSUMOWANIE

Sandberg przedstawił w pracach [6], [7] i [8] inne niż Boyd i Chua podejście do problemu reprezentacji dyskretnego operatora liniowego za pomocą sumy splotowej. W szczególności pokazał [6], że jeżeli dyskretny operator liniowy $G$ jest operatorem ciągłym przekształcającym zbiór sekwencji ograniczonych w zbiór sekwencji ograniczonych, to może on być przedstawiony w następującej postaci:

$$Gx(k) = \sum_{n=-\infty}^{\infty} h(k,n)x(n) + \lim_{m\to\infty}\left(GE_m x(k)\right), \tag{76}$$

gdzie $h(k,n)$ oznacza odpowiedź układu (systemu) opisywanego operatorem $G$ na impuls jednostkowy (jednostkową „próbkę" w postaci delty Kroneckera), przyłożony do tego układu (systemu) w chwili czasowej $k = n$, tj. $h(k,n) = G\delta_{k,n}$. Natomiast funkcja $E_m x(k)$ w (76) jest określona wzorem

$$E_m x(k) = \begin{cases} x(k) & \text{dla } |k| > m \\ 0 & \text{dla } |k| \le m \end{cases} \tag{77}$$

Jeżeli operator $G$ występujący w (76) jest dodatkowo operatorem niezależnym od czasu (stacjonarnym), to można skorzystać z zależności (53), co pozwala napisać

$$h(k,n) = G\delta_{k,n} = G\delta_{k-n,0} = h(k-n,0) \overset{df}{=} h(k-n). \tag{78}$$

Podstawiając w (76) $h(k,n)$ dane wzorem (78) dla operatorów niezależnych od czasu, otrzymujemy w tym przypadku

$$Gx(k) = \sum_{n=-\infty}^{\infty} h(k-n)x(n) + \lim_{m\to\infty}\left(GE_m x(k)\right). \tag{79}$$

Wprowadzając następnie w (79) zamiast zmiennej $n$ nową zmienną $k-n=i$, mamy

$$Gx(k) = \sum_{i=\infty}^{-\infty} h(i)x(k-i) + \lim_{m\to\infty}\left(GE_m x(k)\right) = \sum_{i=-\infty}^{\infty} h(i)x(k-i) + \lim_{m\to\infty}\left(GE_m x(k)\right) \tag{80}$$

Dokonując formalnego podstawienia $i = n$ w (80), otrzymujemy ostatecznie

$$Gx(k) = \sum_{i=-\infty}^{\infty} h(n)x(k-n) + \lim_{m\to\infty}\left(GE_m x(k)\right). \tag{81}$$

Z powyższego wynika, że liniowy, niezależny od czasu oraz ciągły operator $G$, przekształcający zbiór sekwencji ograniczonych w zbiór sekwencji ograniczonych, posiada reprezentację w postaci sumy splotowej

$$\sum_{i=-\infty}^{\infty} h(n)x(k-n). \tag{82a}$$

wtedy i tylko wtedy, gdy drugi składnik po prawej stronie równości (81) równa się zero, to znaczy zachodzi

$$\lim_{m\to\infty}\left(GE_m x(k)\right) = 0. \tag{82b}$$

W kontekście wyników uzyskanych przez Boyda i Chua widać, że zależność (82b) pełni w podejściu Sandberga rolę własności zanikającej pamięci, która to miała decydujące znaczenie w otrzymaniu rezultatu będącego przedmiotem twierdzenia z rozdziału 3. Zauważmy, że do jego osiągnięcia niewystarczający byłby [1] nieco słabszy warunek taki, że operator $G$ jest operatorem skutkowo-przyczynowym (definicja tej własności została podana w rozdz. 2).

Rola zależności (82b) i jej związek z własnością zanikającej pamięci będą przedmiotem dalszych badań.

## BIBLIOGRAFIA

[1] Boyd S., Chua L.O., 1985. Fading memory and the problem of approximating nonlinear operators with Volterra series. IEEE Trans. on Circuits and Systems, vol. CAS-32, pp. 1150-1161.

[2] Barrett J.F., 1963. The use of functionals in the analysis of nonlinear physical systems. Int. J. Electron. Control, vol. 15, pp. 567-615.

[3] George D., 1959. Continuous Nonlinear Systems. MIT RLE Tech. Rep., no. 355.

[4] Volterra V., 1959. Theory of Functionals and of Integral and Integro-Differential Equations. Dover New York.

[5] Wiener N., 1958. Nonlinear Problems in Random Theory. MIT Press Cambridge.

[6] Sandberg I.W., 1998. A note on representation theorems for linear discrete-space systems. Journal of Circuits, Systems and Signal Processing, vol. 17, pp. 703-708.

[7] Sandberg I.W., 2001. A note on the convolution scandal. IEEE Signal Processing Letters, vol. 8, pp. 210-211.

[8] Sandberg I.W., 2006. A short survey of recent representation results for linear system maps. Zeszyty Naukowe ATR seria Telekomunikacja i Elektronika, vol. 9, pp. 9-24.

[9] Borys A., 1997. O aproksymacji nieliniowych systemów cyfrowych. Kwartalnik Elektroniki i Telekomunikacji, z. 1, str. 37-56.

[10] Kuratowski K., 1967. Rachunek różniczkowy i całkowy, funkcje jednej zmiennej. PWN Warszawa, str. 44.

## ON THE DESCRIPTION OF LINEAR DIGITAL CIRCUITS AND SYSTEMS IN THE FORM OF A CONVOLUTION SUM AND THE FADING MEMORY PROPERTY

### Summary

Many physical circuits and systems, analog and digital ones, possess the property of fading memory. For the first time its precise definition has been given by Boyd and Chua. They have also proved that the description of linear digital circuits and systems by the convolution sum is possible if and only if these circuits and systems have the property mentioned above. In this paper, the details of the proof by Boyd and Chua, which were omitted in their original paper, are discussed. These details are important in understanding of another approach to the problem of description of conditions under which the convolution sum representation does exist or not. This new approach was proposed by Sandberg.

Keywords: digital linear circuits and systems, property of fading memory, convolution sum description.

ISSN 1899-0088