EYE AND LIP ENHANCEMENT FOR LOWRATE VIDEOPHONY

J. Streit, L. Hanzo, B. Bradshaw

Dept. of Electr. and Comp. Sc., Univ. of Southampton, SO17 1BJ, UK.
Tel: +44 1703 593 125, Fax: +44 1703 593 045
Email: jss@ecs.soton.ac.uk, lh@ecs.soton.ac.uk
WWW: http://www-mobile.ecs.soton.ac.uk

ABSTRACT

A generic parametric coding enhancement for employment in very low-rate videophone systems is proposed in order to improve the subjectively important eye and lip representation. The four-step Algorithm 1 identifies the position of the eyes, nostrils, lips and nose using pre-precessing and pattern-matching and after luminance adjustment and smoothing it overlays the best-matching entries from a trained codebook, if deemed beneficial in image quality terms. The benefits of the technique are demonstrated using the example of an 11.36 kbps fixed-rate quad-tree codec.

1. MODEL-BASED PARAMETRIC ENHANCEMENT

Although in low-rate videophony the high-quality representation of the subjectively most important eye and lip regions is paramount [1], due to the paucity of bits imposed by the system's bandlimitation this cannot be always automatically ensured. Explicitly, when using 1000 bits/frame for the encoding of Quarter Common Intermediate Format (QCIF) images, as in Reference [5], only too small a fraction of the 25 384-pixel frame could be adequately modelled. Hence in these highly active spots we propose to employ model-based parametric coding (PC) enhancement. The penalty associated with this technique is, however that the decoder needs to store the eye and lip codebooks. Alternatively, these codebooks must be transmitted to the decoder during the call setup phase. Note however that otherwise the proposed technique is general and it can be invoked in conjunction with any coding technique.

Parameterisation of images grew in popularity during the 1980's along with the improvement in computer graphics techniques. In 1982 Parke et al [3] investigated the realism of generated human faces relying on manual measurement of facial profiles and hence the results of this approach were somewhat artificial. Yau and Duffy [2] proposed a system whereby a HeNe laser was used to generate a 'wire-frame' three-dimensional (3D) model. Image reconstruction was achieved by pasting an image of the subject's face over the underlying polygon model. By successfully generating this structure, rotation, translation and scaling could easily be performed. Welsh [4] also worked extensively in this research area and suggested a variety of techniques for encoding both eye- and mouth-feature codebooks as well as 3D models.

During our following discourse Sections 2 and 3 describe our approach to the eye and mouth detection and the parametric database training, while Section 4 highlights the encoding process. Lastly, Section 6 characterises the video quality enhancement of the proposed scheme in the context of a previously proposed 11.36 kbps Quad-tree (QT) video codec. Let us commence with the description of the proposed scheme, before we highlight, how it can be invoked in a videophone codec.

2. EYE AND MOUTH DETECTION

Eye and mouth detection has been the aim of various research interests such as model based coding, vision assisted speech recognition [6] and lip-reading [7]. A reliable detection of the eye and mouth location is the most crucial step for all of the above techniques.

Step 1: It is critical for the reliable operation of the proposed eye-mouth detection algorithm that the face is the largest symmetrical object in the head-and-shoulders image frame and that it's axis of symmetry is vertical. Initially we generate a black and white two-tone image from the incoming frame in order to detect this symmetry and to simplify the detection process. This two-tone image is free from gradual brightness changes but retains all edges and object borders. This was achieved by smoothing the picture using a simple two dimensional averaging FIR filter and then subjecting the smoothed image to frame differencing and thresholding. Our experiments were based on 3×3 and 5×5 order filters, where all 9 or 25 filter coefficients were set to 1/9 and 1/25, respectively, in order to preserve energy. The filtered image was then subtracted from the incoming frame and finally thresholded, which led to a binary image $f_{bin}(x,y)$ similar to the one shown at the left hand side Figure 1. A threshold of 8.0 has been found suitable for this operation. This concludes the first step of the parametric coding algorithm summarised in Algorithm 1.

Step 2: The next step is to find the axis of symmetry for the face. Assuming tentatively that the axis of symmetry is x_0, symmetry to this axis is tested by counting the number of symmetric pixels in the two-tone image. The specific x_0 value yielding the highest number of symmetric pixels is then deemed to be the axis of symmetry. Once this axis is known, the pixel-symmetric two-tone image $f_{sym}(x,y)$ at the right hand side (RHS) of Figure 1 is generated from the

PIMRC'96, TAPEI, TAIWAN

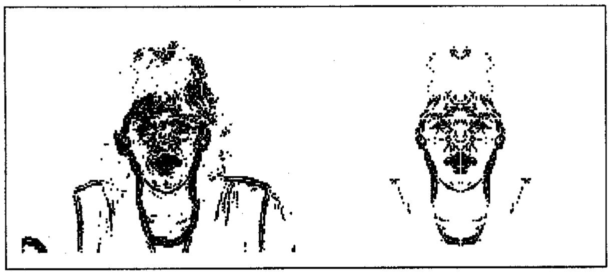

Figure 1: Binary (LHS) and binary-symmetric (RHS) frame of the 'Missa' sequence generated by Steps 1 and 2 of Algorithm 1

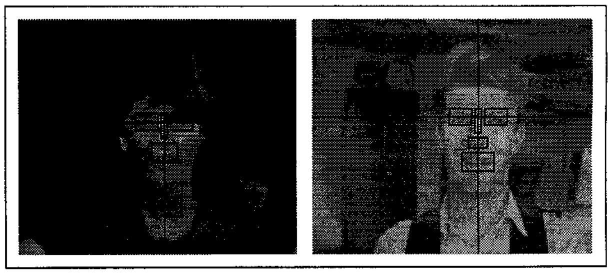

Figure 2: Image frames with overlaid initial (left) and improved (right) templates used in Algorithm 1

binary image $f_{bin}(x,y)$ using Equation 1:

$$f_{sym}(x,y) = f_{bin}(2x_0 - x, y) f_{bin}(x,y), \quad (1)$$

which simply eliminates the non-symmetric pixels from the two-tone image.

Algorithm 1 *This algorithm summarises the parametric coding enhancement steps.*

```
1. Generate the binary image using smoothing, frame differencing and thresholding (Figure 1).
2. Identify the axis of symmetry yielding the maximum number of symmetric pixels (Equation 1).
3. Identify the position of eyes, nostrils, lips and nose using the scaled template (Figure 2).
4. Find the best matching eye and lip codebook entries and send their position, luminance shift and codebook index.
```

Step 3: The pixel-symmetric image is then used to localise the eyes and the mouth. Initially we attempted to locate the eye and the mouth as the two most symmetrical objects, assuming a given axis of symmetry x_0 in the frame $f_{sym}(x,y)$. This technique resulted in a detection probability of around 80 % for the 'Miss America' and 'Lab' sequences. An object was deemed to be correctly detected if it's true location in the original frame was detected with a precision of +/- 4 pixels. In most of the cases the location of the eyes was correctly identified, but often the chin was erroneously detected by the Algorithm as the mouth.

Hence in a refined approach we contrived a more sophisticated template, which consisted of separate areas for the eyes, the nose and a combined area for the nostrils and the mouth, as seen at the LHS of Figure 2. We expect many 'contrast pixels' in the binary symmetric image $f_{sym}(x,y)$ at the location of the eyes, the mouth and nostrils, while the nose is not conspicuously represented in Figure 1. The template was scalable in the range of 0.8 to 1.1 in order to cater for a range of face sizes and distances measured from the camera. This template was then vertically slid along the previously determined axis of symmetry and at each position the number of symmetric pairs of contrast pixels appearing within the template overlayed on the binary symmetric image $f_{sym}(x,y)$ was determined. An exception was constituted by the nose rectangle, where the original luminance change was expected to be gradual, leading to no contrast pixels at all. This premise was amalgamated with our identification procedure by reducing the total number of 'symmetric contrast pixels' detected within the confines of the template by the number of such contrast pixel found within the nose rectangle. Clearly, a high number of 'symmetric contrast pixels' in the current presumed nose region weakened the confidence that the current template position was associated with the true position. Finally, the template location resulting in the highest number of matching symmetric pixels was deemed to be the true position of the template.

Although due to these measures the eye and lip detection probability was increased by another 10 %, in some cases the nostrils were mistaken for the mouth. We further improved the template by adding a separate rectangular area for the nostrils as depicted at the RHS of Figure 2. This method reached a detection probability of 97 % for our test sequences which was deemed adequate for our purposes.

In our previous endeavours the eye and lip detection technique outlined was tested using image sequences, where the speaker keeps a constant distance from the camera and hence the size of the face remains unchanged. However, the algorithm can adapt to the more realistic situation of encountering a time-variant face size by scaling the template. The magnification of the template was allowed to vary in the range of 0.8 to 1.1. The algorithm was tested using various 'head and shoulder' sequences and we found that it performed well as long as the speaker was sufficiently close to the camera. Even the fact that a speaker wore glasses did not degrade the detection performance, although difficulties appeared in case of individuals wearing a beard. In these situations the parametric QT codec enhancement had to be disabled on grounds of reduced mean squared error (MSE) performance, which was signalled to the decoder using a one-bit flag. Again, the above parametric coding steps are summarised in Algorithm 1 and Step 4 will be elaborated on during our further discussions.

3. PARAMETRIC CODEBOOK TRAINING

A critical issue as regards to the parametric codec's subjective performance is the training of the eye and lip codebooks. Large codebooks have better performance and higher complexity than small ones. The previously described eye and lip identification algorithm can be invoked to train the codebooks, which can then be manually edited in order to remove redundant entries and hence reduce the codebook search complexity. Initially we generated a single codebook for the eyes and derived the second eye-codebook by mirroring the captured codebook entries. As this could lead to

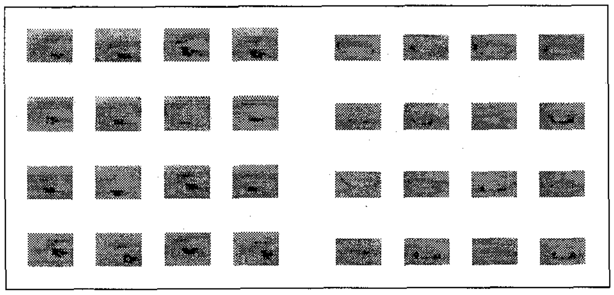

Figure 3: Eye and lip codebooks for the parametric coding of the 'Lab Sequence'

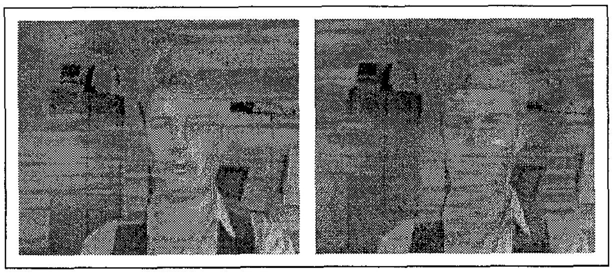

Figure 4: Parametric coding example: The 'Lab' sequence coded with entries from the 'Miss America' codebook

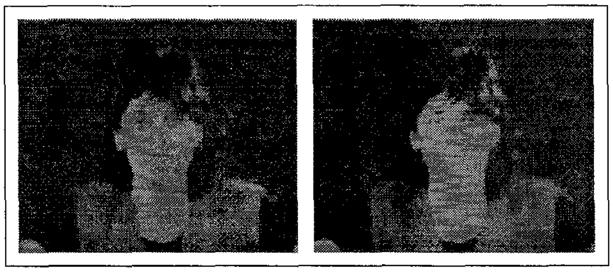

Figure 5: Subjective video quality of the QT codec without (left) and with (right) PC at a bit rate of 11.36 kbps or 1136 bits/fr

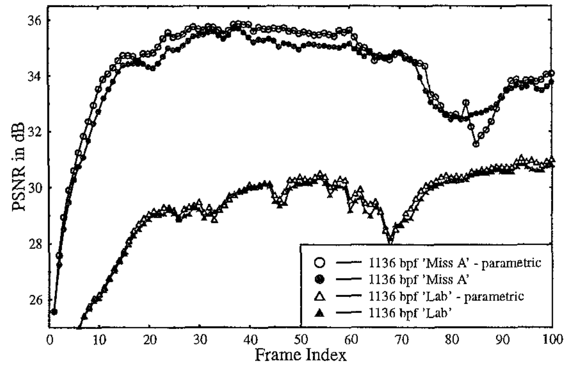

Figure 6: PSNR versus frame index performance for the QT codec at 1136 bits/frame with and without PC

pattern matching problems, when the head was inclined, we proceeded with separate codebooks for each eye. A sample codebook of 16 eye- and lip-entries for the 'Lab' sequence is portrayed in Figure 3, where the contrast of the entries was enhanced for viewing convenience.

4. PARAMETRIC ENCODING

Step 4: The parametric encoder detects the eyes and mouth using the template matching approach described above. Once the template position is detected the codec attempts to fit every codebook entry in to the appropriate locations by sliding it over a window of +/- 3 pixels in each spatial direction. The MSE for each matching attempt is compared and the best matching entry is chosen. A luminance shifting operation allows the codec to appropriately adjust the brightness of the codebook entries. This is necessary as the brightness during codebook generation and encoding may differ. An example of the subjective effects of parametric coding and enhancement (PC) is portrayed in Figure 4 where we attempted to match the eye and mouth entries of the 'Missa' sequence into the 'Lab' sequence in order to assess the codec's performance outside the training sequence. This scenario would represent the worst case situation, when a prestored parametric codebook is used, which does not contain entries from the current user, instead of training the decoder's codebook during the call setup phase. Observe that there are no annoying artifacts, although the character of the face appears somewhat different. Hence, if the system protocols allow, it is preferable to opt for codebook training before the commencement of the communication phase.

The above optional parametric coding (PC) steps were included in our 11.36 kbps QT codec of Reference [5]. If PC affected the video quality advantageously in reconstructed mean-squared error terms, it was enabled, otherwise disabled. If the parametric enhancement was deemed to be successful for a give frame, then for each modelled object the exact position, luminance shift and codebook index was transmitted. This required, including the one bit PC enable flag, a total of 70 bits as detailed in Table 1. Explicitly, assuming arbitrary, independent eye and lip locations, for each object a 15-bit position identifier is required, yielding a total of 45 bits. When using 16 independent luminance shifts for the three objects a total of 12 bits are necessary, similarly to the object codebook indices. This 70 bit segment contained some residual redundancy, since the object positions and luminance shifts are correlated to both each other and to their counterparts in consecutive frames, which would allow us to reduce the above 45-bit location identifier to around 32 bits and the total number of bits to about 57, but this further compression potential was not exploited here. Let us now briefly demonstrate, how the proposed eye and lip enhencement can be invoked in lowrate video codecs, using the example of quad-tree coding. We note however that the technique is applicable to arbitrary low-rate codecs.

5. QUAD-TREE CODEC

Our proposed quad-tree (QT) based codec was portrayed in depth in Reference [5], hence here only a rudimentary decription is offered. The block diagram of the proposed QT codec processing 176 × 144 pixels Quarter Common Intermediate Format (QCIF) images scanned at 10 frames/s

Type	Location of the object	Luminance shift	Codebook entry	PC flag	Total
Bits required	3×15 = 45	3×4 = 12	3×4 = 12	1	69

Table 1: Bit allocation for the parametric codec enhancement

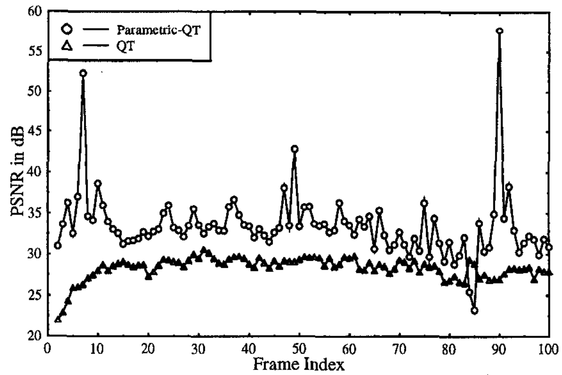

Figure 7: PSNR versus frame index performance for the parametrically enhanced QT codec in the template region using the Lab Sequence

is shown in Figure 8. For videophony over conventional mobile radio speech channels, such as the Pan-European GSM or the American IS-54 and IS-95 systems, fixed-rate video codecs are required. As it is seen at the top of the codec's schematic, the intra-frame coding mode is invoked at the commencement of communications, during which a low-resolution initial image is transmitted to the decoder in order assist in its start-up phase, as it will be described shortly. In the motion prediction (MP) block of Figure 8 the QCIF frame is first segmented into small, for example 8×8-pixel perfectly tiling blocks. Then each block is slid over a certain motion-velocity and frame-scanning rate dependent search area of the previous reconstructed frame and it is estimated by finding the position of highest correlation, which location each block was deemed to have originated from due to motion translation. The corresponding coordinates referred to as motion vectors (MV) are then used in the motion compensation (MC) process of Figure 8 to appropriately position each incoming block, which are then subtracted from the previously reconstructed frame in order to generate the so-called motion compensated error residual (MCER). The MCER is QT-decomposed to the required resolution under the joint action of the Bitrate Control and Tree Reduction blocks. In the reconstruction process the MVs assist at both the local and the distant decoder in an inverse fashion in order to appropriately update the reconstruction frame buffers. The eye and lip representation quality of the QT codec can be improved by the optional Parametric Coding (PC) arrangement of Figure 8.

Due to transmission errors the encoder's and decoder's previous reconstructed frame buffers may become misaligned, which leads to prolonged artifacts at the output of the decoder. This effect can be remedied using Partial Forced Updates (PFU) in order to identically replenish the encoder's and decoder's previous reconstructed frame buffers. In our proposed codecs we were constrained to a simple PFU technique using the down-scaled and partially overlaid 4-bit encoded 8×8-bit block averages, which was applied only to a fraction of the blocks in each frame due to the tight bit rate budget available. Although the PFU process partially overwrites the previous reconstructed buffers at both the local and remote decoder by the above mentioned very crude image estimate, therefore slightly impairing the codec's error-free performance, in case of high channel bit error rates (BER) it has an error mitigating effect by gradually replenishing both buffers' contents.

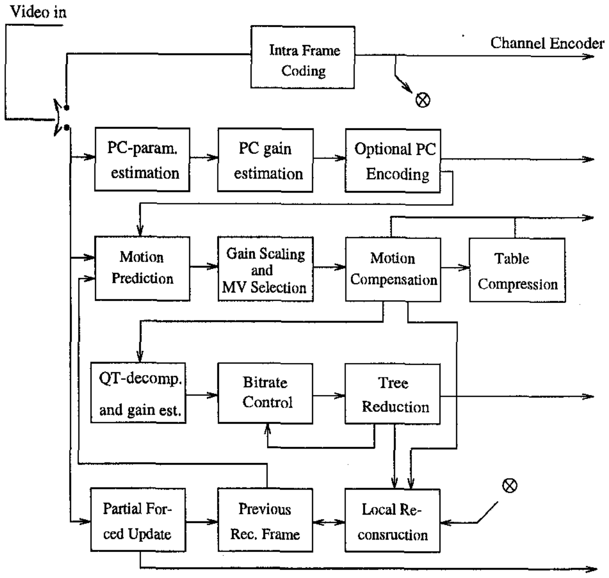

Figure 8: QT Codec Schematic

The number of 8×8-pixel PFU blocks per QCIF frame depends on the target bit rate and it is automatically determined by the proposed re-configurable codec. Consequently the frequency of partially forced updating a certain block is also bit rate dependent, although higher prevailing BER values would require more frequent updates, irrespective of the bit rate budget. In our 11.36 kbps QT videophone codec 20 randomly scattered 8×8 blocks out of the 396 blocks per frame are updated, which requires 80 bits/frame. This implies that the PFU frequency of each specific block is about 1/(2 s), which is equivalent to updating the same block about every 2 s or 20 frames. During PFU both the local and remote reconstructed buffers' contents are scaled by 0.7 and the 0.3-scaled 4-bit quantised block averages are superimposed, allowing a non-destructive gradual replenishment to take place.

The motion compensation (MC) scheme determines a 4-bit motion vector (MV) for each of the 8×8 blocks within a search window of 4 × 4 = 16 pixels using full-search. The potential gain of MC is assessed in terms of Motion Compensated Error Residual (MCER) energy reduction and the gains in the subjectively important eye and mouth region may be augmented by a factor of two. A bit-rate dependent number of 'motion-active' blocks is then subjected to full motion compensation, while for the 'motion-passive' blocks frame differencing is employed. Each of the 396 blocks would require a 9-bit identifier, leading to a total of $9 + 4 = 13$ bits per active vector. Typical motion activity rates around 60 MV would use most of the available bit rate budget. In order to accommodate around 60 active MVs within a budget of 500 bits, we assign a 1-bit motion-activity flag for each vector and hence create a MV activity table of 396 bits. This motion activity table can be compressed by about a factor of two using the run-length based 'Table Compression' algorithm of Figure 8, when aiming for a target bit rate budget of around 1000 bits/frame or 10 kbps [5].

6. CODEC PERFORMANCE

The performance of the proposed parametric coding enhancement was tested in the context of the above 11.36 kbps, 10 frames/s scanned, parametrically enhanced fixed-rate QT codec [5] [1]. The bit allocation scheme of the codec is summarised in Table 2. From the bitrate budget of 1136 bits/frame a total of less than 500 bits/frame were dedicated to the motion compensation activity table and the motion vectors, which were allocated using an intelligent cost-gain quantised approach [5]. Furthermore, 1 or 70 bits were earmarked for parametric coding, 80 bits for partial forced updates (PFU) in order to mitigate the effect of transmission errors, while the remaining 486 or 567 bits were dynamically assigned to variable-length QT coding using the adaptive bitallocation scheme of Algorithm 2, which was decribed in depth in Reference [5]. Suffice to say here that this Algorithm allowed us to eliminate the specific leaves from the QT that resulted in the lowest decomposition gains, while providing a powerful cost-gain quantised bit rate control protocol.

The objective Peak Signal to Noise Ratio (PSNR) versus frame index performance of this codec is portrayed in Figure 6 both with and without parametric enhancement. Observe in the Figure that the PSNR curve of the Miss America sequence temporarily caves in between frames 75 and 85 due to the fixed bit rate limitation, as she moves her head rather abruptly. Furthermore, due to the relatively small, 470-pixel area of the eye and lip regions and because of the optional employment of the parametric enhancement, the overall objective PSNR video quality improvement appears more limited than its subjective improvements exemplified in Figure 5. Explicitly, the objective PSNR improvement due to the parametric enhancement is de-weighted by the factor of the template-to-frame area ratio, namely by $470/(176 \times 144) = 0.0185$. The PSNR improvement of the template area becomes more explicit in Figure 7, which is the equivalent of Figure 6 related to this smaller 470-pixel image frame section. Observe in Figure 6 that for the 'Miss America' sequence PSNR values in excess of 34 dB are possible at 11.36 kbps. The proposed scheme can be invoked advantageously in arbitrary low-rate codecs.

Algorithm 2 *This algorithm adaptively adjusts the required QT resolution, the number of QT description bits and the number of encoding bits required in order to arrive at the target bit rate.*

```
1. Develop the full tree from minimum to
   maximum number of QT levels (eg 2-7).
2. Determine the MSE gains associated with all
   decomposition steps for the full QT.
3. Determine the average decomposition gain
   over the full set of leaves.
4. If the potentially required number of coding
   bits is more than twice the target number of
   bits for the frame, then delete all leaves
   having less than average gains and repeat
   Step 3.
5. Otherwise delete leaves on an individual
   basis, starting with the lowest gain leaf,
   until the required number of bits is
   attained.
```

Parameter	MC	PC	PFU	QT	Total
No. of bits	< 500	1 or 70	80	486 or 567	1136

Table 2: Bit allocation for the 11.36 kbps Codec 1 using optional PC enhancement

7. ACKNOWLEDGEMENT

The financial support of the EPSRC, UK in the framework of the research contract GR/K74043 is gratefully acknowledged.

8. REFERENCES

[1] **D. Morris**: Manwatching, Jonathan Cape, London, 1977

[2] **Duffy N.D. and Yau J.F.S.**: Facial image reconstruction and manipulation from measurements obtained using a structured lighting technique, Pattern Recognition Letters, April 1988, Vol. 7, pp239-243

[3] **Parke F.I.**: Parameterised models for facial animation, IEEE Computer Graphics and Applications, November 1982, Vol. 12, pp 61-68

[4] **Welsh W.J. and Shah D.**: Facial feature Image coding using Principle Components, Electonics Letters, October 1992, Vol. 28, pp 2066-2067.

[5] **J. Streit, L. Hanzo**: Quadtree-based parametric wireless videophone systems, IEEE Tr. on Circuits and Systems for Video Technology, Special Issue on Wireless Visual Communications, Vol. 6, No. 2, Apr. 1996, pp 225-237

[6] **M. Hennecke et al**: Using Deformable Templates to Infer Visual Speech Dynamics, 28th Annual Asilomar Conference On Signals, Systems, and Computers, 1994

[7] **G.J. Wolf et al.**: Lipreading by neural networks: Visual preprocessing, learning and sensory integration, Proceedings of the neural information processing systems, Vol 6, 1994, pp 1027–1034

[1] The coding performance of this codec along with a range of other associated schemes can be assessed with reference to the WWW homepage http://www-mobile.ecs.soton.ac.uk. A downloadable demonstration package is also available under this address