Lip Synchronization in 3-D Model Based Coding for Video-conferencing

J. A. Provine and L. T. Bruton

***Abstract*—One of the visually important aspects in a video-conferencing sequence is lip synchronization (lip-synch), particularly when a 3-D model based coding scheme is used to transmit the sequence. This issue has not been addressed so far by model based coding schemes. We present a technique to synchronize the lip movements to the audio associated with the sequence for a coding scheme that uses a muscle based facial model.**

I. Introduction

MODEL based coding is an encoding scheme that has received significant attention lately, for transmitting talking head sequences(as in video-conferencing) at very low bit rates[1]-[6]. The underlying principle is to use a model at the transmitting end to identify various actions performed by the subject in the sequence, and transmit only those parameters that are required to *synthesize* the sequence using an identical model. The synthesis here involves the actual creation of a sequence *perceptually* similar to the input sequence. The model used at the synthesizer is texture mapped with the subject's image to produce the desired identity. In this paper, model based coding is used to refer to methods that strictly use a 3-D facial model to encode talking head type sequences, though this term has also been used to refer to object-based coding schemes[3]-[6].

One of the vital issues involved in synthesizing a realistic talking head is the coordination of the lip movements with the associated speech. Lip-synch can be accomplished just as any other facial expression, only here the movements are timed with the associated speech signal. The model parameters necessary to synthesize the lip movements are obtained from the fundamental speech units of the associated audio. Though lip-synch is important for creating a realistic sequence, it has not received any attention in model based coding literature so far. We describe a method to accomplish lip-synch where the model parameters required for synthesizing only the key frames associated with the fundamental speech units need to be transmitted. Inbetweening is performed using the timing information from the associated audio, thereby reducing the number of bits required for synthesizing a lip synchronized talking head. Lip-synch has received considerable attention in the area of character animation[7]-[8], where it is achieved by directly manipulating the vertices of polygons in the surface approximation. The material presented here is also applicable for obtaining lip-synch in character animation using a muscle based model.

The authors are with the Department of Electrical and Computer Engineering, University of Calgary, Calgary, Canada.

II. 3-D muscle based facial models

Muscle based facial models use simulated muscles to indirectly displace the surface, thereby generating the desired facial expressions. The surface could be approximated by polygonal mesh[9]-[10] or bicubic splines[11]. The 3-D facial model referred to here has its external surface approximated by hierarchical splines[11]. Each simulated muscle has its attachment point on the surface and insertion point in the corresponding hypothetical bone. Each of these simulated muscles displace a predefined set of control points on the surface. The direction and the amount of displacement depends on the type of muscle activated. The jaw joint controls the opening of the mouth.

Three different types of muscles are simulated in the model: i) fan, ii) sheet and iii)sphincter. The predefined surface points associated with these three types of muscles are such that the areas of influence are fan-shaped, rectangular and elliptical respectively. The deformations on the surface are handled by displacing the surface points to produce appropriate bulge on the surface[11]. While simulating the muscle's relaxed or contracted posture, care is taken to conserve the volume occupied by the muscle. Facial expressions are sometimes created by the interaction of different muscles with overlapping areas of influence rather than by just a single muscle or jaw action. The displacement of each surface point within a specific region is handled as follows: if $\mathbf{m}_i$ is the vector representing the influence of the i^{th} muscle acting on the surface point, the magnitude of the displacement is governed by $min\{\sum|\mathbf{m}_i|, max(\mathbf{m}_i)\}$ and the direction of the movement of the surface point by the direction of $(\sum \mathbf{m}_i)$[11].

Since the muscles and their areas of influence are predefined in the model, the maximum factor by which the muscle can contract without displacing the surface points to absurd positions is normalized to 1(The muscle in its relaxed state has a contraction factor of 0). Muscles along with appropriate contraction factors are then adequate to create any particular facial expression. By transmitting merely the muscles (or joint) involved and their contraction factors, we avoid directly transmitting the surface points and their displacement, thereby reducing the actual amount of information transmitted to render each scene. This is justified here because, the models at either ends of the codec are identical. With this knowledge of muscle based models, we proceed to describe our lip-synch algorithm. The curious reader is referred to [9] and [11] for more information on muscle based facial models.

0-7803-2570-2/95 $4.00

III. Automated lip-synch in the synthesized sequence

A. Mouth positions corresponding to the fundamental speech units

With the assumption that the speech signal associated with the video-conferencing sequence is in English, the speech signal is decomposed into fundamental speech units, phonemes and diphthongs (from now on, we shall indicate them collectively as FUs). Furthermore, we assume that such units are available at the decoder or the encoder, in which case it is made available to the decoder directly or indirectly (as model parameters). Lists of phonemes and diphthongs used to synthesize lip synchronized talking head are provided in tables 1 and 2 respectively. The FUs can also be used in a speech synthesizer to synthesize the corresponding speech signal[12].

The technique used to produce various mouth positions for the phonemes and diphthongs is similar to lip readers identifying words[13]. Only in our case, we have the phonemes and diphthongs and use them to create the mouth positions. These positions are used as templates to create a lip synchronized sequence corresponding to the audio associated with the sequence. Ekman and Friesnen [14] have devised action units (AU) corresponding to different facial expressions in their facial action coding system (FACS), which is a familiar treatise to researchers in model based coding. The muscles involved in producing different AUs related to a particular FU are activated to obtain the corresponding mouth position.

Mouth positions of a talking head is not only dependent on the FU but its position within a word and a sentence. This would mean that we need to create mouth positions for different allophones (which are the phonemes according to the positions they take within a word). This would mean creating over 200 templates. However, as it will become obvious from our simulation results, it is adequate to produce templates for just 40 FUs. In synthesizing the actual frames, the synthesizer *looks ahead* for the contraction parameters of the muscles involved in creating the mouth position corresponding to the subsequent FU. The sequence is created with a smooth change from the current position to the next position using *parametric interpolation* described in Section III B. Homophones do not present a problem as in speech recognition because, the mouth positions look alike to the viewer.

The reason for distinguishing the phonemes from diphthongs is because of the number of mouth positions associated with each of these units. Phonemes have only one mouth position associated with them, while diphthongs have two positions, a short initial position and a longer final position. Hence, the synthesizer should use the appropriate muscle contraction parameters to simulate both positions within the duration of the diphthong. The emphasis on this fact is easily identified by considering the following simple example: Consider the words "cow" and "coup". The former has a diphthong /OW/ following /k/. If we associate only one mouth position to this diphthong, the lip movements corresponding to both these words are identical, which is incorrect.

Phoneme	Muscle/ Joint	Pull/ Rotation Factor	Example
AA	Jaw	94.25	not(nAAt)
AE	ZMJ Jaw	0.85 94	bat (bAEt)
AH	Jaw	94.5	but(bAHt)
AO	ILI Jaw	0.75 95	soft (sAOft)
AX	Jaw	96	bottle (bAAttAXl)
CH	ILI&S Jaw	1 92	church (CHERCH)
DH	Jaw	90.75	this(DHIHs)
EH	ZMJ Jaw	0.65 92	let (lEHt)
ER	Jaw	92	bird(bERd)
IH	Jaw	93.5	fit(fIHt)
IY	ZMJ Jaw	0.9 91.5	speak (spIYk)
SH	ILI Jaw	1.5 93	shine (SHAYn)
TH	Jaw	90.75	thaw(THAO)
UH	ILI&S Jaw	1 90.4	book (bUHk)
UW	ILI&S Jaw	1 90.4	to (tUW)
WH	ILI&S Jaw	1 91	when (WHEHn)
ZH	ILI Jaw	1.5 93	measure (mEHZHER)
d	Jaw	90.75	date(dEYt)
f	ILI&S	1	fence (fEHns)
g	Jaw	91.5	get(gEHt)
j	ILI&S Jaw	1 92.5	gem (jEHm)
k	Jaw	91.5	cute(kyUWt)
l	ILI Jaw	1 96	light (lAYt)
n	Jaw	91.5	now(nAW)
r	ILI&S Jaw	0.5 90.5	rare (rEHr)
s	ZMJ Jaw	0.5 90.75	speak (spIYk)
t	Jaw	91.5	to(tUW)
v	ILI Mentalis Jaw	1 1 90.75	van (vAEn)
w	ILI Jaw	1 95.5	wet (wEHt)
y	Jaw	91.5	yes(yEHs)
z	Jaw	90.75	zoom(zUWm)

Table 1. Defaults to create mouth positions for phonemes.

The following 6 pairs of muscles (all of these muscles on one side of the face are reflected to simulate muscles on the other side of the face) can be used to produce the mouth positions for the FUs: i) Zygomatic major (ZMJ) (pull the corners of the lips), ii) Incisivus Labii Superioris (ILS) (pucker the lips), iii) Incisivus Labii Inferioris (ILI) (pucker the lips), iv) Mentalis (raise the chin), v) Risoris (pull the corner of the lips back) and vi) Orbicularis Oris (funnel, tighten or press the lips). In addition to these muscles, the jaw joint is rotated to control the mouth opening corresponding to each FU. Anatomically, it is not as simple as using a subset of these muscles and joint to produce the mouth position for each FU. However, an output of

Diphthong	Muscle/ Joint	Pull/ Rotation Factor	Example
AW			now(nAW)
1	Jaw	94	
2	ILI&S	1	
	Jaw	91	
AY			time(tAYm)
1	Jaw	93.5	
2	Jaw	91.4	
EY			bait(bEYt)
1	Jaw	91.25	
2	ZMJ	0.9	
	Jaw	91.25	
OW			note(nOWt)
1	ILI	0.5	
	Jaw	93	
2	ILI	1	
	Jaw	96	
OY			toy(tOY)
1	ILI	0.5	
	Jaw	94	
2	ILI	0.25	
	Jaw	90.75	

Table 2. Defaults to create mouth positions for diphthongs.

perceptually good quality can be created using these muscles and joint. The contraction parameters corresponding to the muscles and the joint involved to produce the positions for each FU are given in tables 1 and 2. The absence of a value for a particular muscle denotes that it is in its relaxed state(when the mouth is closed, the jaw has a rotation angle = 90°). The phonemes excluded in table 1 are /p/, /b/, /m/ and /h/. For the first three phonemes, the above six pairs of muscles are relaxed and the mouth is closed. /h/ is usually identified with the chest heaving and the mouth position corresponds to that of the subsequent FU. A phoneme that needs special mention is /NG/. For this phoneme, the previous position is held with the sound being nasal. Other than the muscles and the joint the tongue and teeth influence the way certain phonemes are pronounced as in /f/ and /DH/.

During a typical video-conferencing sequence, there are other facial expressions that occur apart from those connected with speech. Some such expressions are produced by some of the muscles involved in lip-synch. In such cases the parameter from the analyzer should compensate for the involvement of a particular muscle in two facial expressions simultaneously. However, to be careful not to produce absurd surface displacement, if the muscle pull factor for a particular expression is c_k and the factor for producing the position corresponding to the FU at that instant is c_l, the synthesizer uses $max\{c_k, c_l\}$ to pull that muscle while reproducing the scene.

B. Coordinating output sequence generation to the audio

The muscle and jaw parameters provided in tables 1 and 2 are used by the synthesizer to produce the key frames corresponding to the FUs. However, we need to perform inbetweening to generate an output sequence that contains smooth changes between successive positions. Fortunately, the muscle and jaw parameters involved in creating facial expressions lend themselves to *parametric interpolation* as opposed to interpolating pixel by pixel or block by block. This approach reduces the amount of information need to be transmitted as well as the time required to generate the frames. The parameter in question is the muscle pull factor or jaw rotation angle. Two cases, i) interpolating *from* the position for phoneme or final position for diphthong *to* position for the next phoneme or initial position for a diphthong and ii) interpolating *from* initial position of a diphthong *to* a phoneme or initial position of another diphthong are handled separately.

Let the duration for each FU be t sec. Then, for a sequence synthesized at the rate of R frames/sec. (R= 30 for NTSC and 25 for PAL, SECAM), n, the number of frames between the initial and final key frames is given by

$$n = \begin{cases} \lfloor Rt \rfloor & \text{if } \lfloor Rt \rfloor > k \quad \text{and} \\ 0 & \text{otherwise,} \end{cases} \tag{1}$$

where $\lfloor x \rfloor$ denotes the smallest integer $\leq x$, $k = 1$ for phoneme and $k = 2$ for diphthong. Note that the value n does not include the key frames themselves. *Case (i)*: The muscle contraction parameter or the jaw rotation angle of an active muscle/joint, used to create frame i, is

$$c_i = min\{c_f, c_l\} + \frac{j\,|c_f - c_l|}{n}, \quad \text{for } 1 \leq i < n, \tag{2}$$

where $j = (n - i)$ if $c_f > c_l$, $j = i$ if $c_f \leq c_l$ and c_f and c_l are the factors for the two successive key frames involved. *Case (ii)*: The total number of interpolated frames, n calculated as specified in equation (1) is split into $n_1 = \lfloor \frac{n}{3} \rfloor$ and $n_2 = (n - n_1)$, with n_1 representing the number of frames between the first position and the second, and n_2 representing the number of frames between the second position and the position for the succeeding FU. The parameters used to create the frames are calculated as follows:

$$c_i = min\{c_f, c_m\} + \frac{j\,|c_f - c_m|}{n_1}, \quad \text{for } 1 \leq i < n_1, \tag{3}$$

where $j = i$ if $c_m \geq c_f$, $j = n_1 - i$ if $c_m < c_f$ and c_m is the factor corresponding to the second position for a diphthong, and

$$c_i = min\{c_m, c_l\} + \frac{j\,|c_m - c_l|}{n_2}, \quad \text{for } 1 \leq i \leq n_2, \tag{4}$$

where $j = i$ if $c_l \geq c_m$ and $j = n_2 - i$ if $c_l < c_m$. These calculations are made for all muscles and joint involved in the creation of a new scene. The use of the same linear function used to interpolate all the muscles and joint involved is justified, because these calculations are transparent to the manner in which the surface points within the region of influence are displaced for different muscles. Note that this process is independent of what takes place in other regions of the face. For instance, lateral frontalis can be suitably contracted to indicate a surprised look while the appropriate muscles and jaw are used to produce the desired mouth position. When one of the muscles involved in synthesizing the mouth position is also used in a different facial expression, the parameter for that muscle is adjusted as explained in Section III A.

C. Simulation Results

A lip synchronized talking head sequence was synthesized using the procedure described above. Though the actual test would be to assess the sequence generated, we provide the the key frames of a simple word "now", which contains a phoneme /n/ and a diphthong /AW/ in it. The mouth positions of the model (picture on the left) are compared against that of a real person uttering the same word (picture on the right) in Fig. 1 - 3. The positions indicate a perceptually close match.

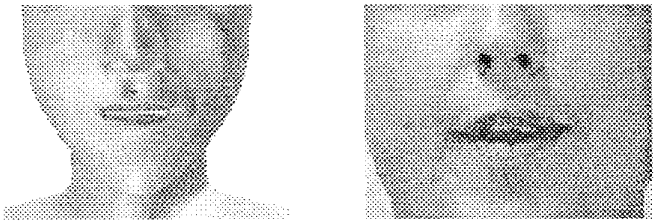

Fig. 1. Positions for phoneme /n/

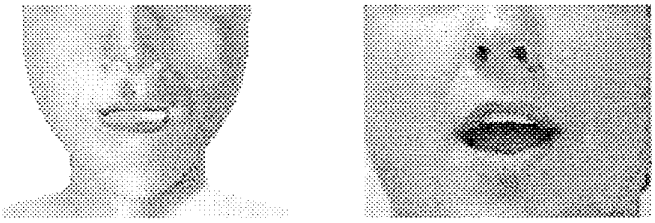

Fig. 2. Initial Positions for diphthong /AW/

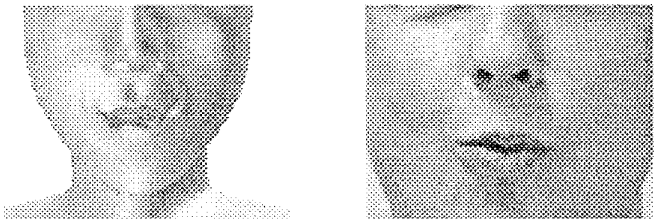

Fig. 3. Final positions for diphthong /AW/

The estimated number of bits dedicated to lip-synch for every key frame is very low. In the worst case scenario, there will be six muscle pairs and a joint activated each with a factor represented by B bits, like any other facial muscle. B depends on the model's sensitivity to the precision of the contraction parameter (or rotation angle). For instance, $B = 16$ is a good choice. If there are a total of N muscles available in the model, the total number of bits required when all the six muscles and joint are activated is $13(\lceil \log_2(N+1) \rceil + B)$ bits. This is not an overhead to the number of bits required to synthesize the sequence, but a part of it. The number of bits represented here is for the binary representation of the parameters before any further encoding. Also, the resolution of the image is dependent on the capability of the synthesizer, and does not affect the number of bits actually transmitted to reproduce the scene.

IV. Conclusions

Lip-synch is an important issue in transmitting a talking head sequence, as in video-conferencing, using a model based codec. Muscles and the jaw joint that are used to produce any facial expression are used in conjunction with the FUs obtained from the speech to obtain lip-synch. Parametric interpolation is used to generate the sequence at the desired frame rate. Using this approach, an output comparable to a real talking head can be produced. The scheme is neither bit expensive, nor resolution dependent. The method is also independent of surface approximation, and hence is equally applicable to a polygon-mesh model. Furthermore, it can also be used in other applications such as character animation.

Acknowledgments

The authors thank Carol Wang for her help with the facial animation system and related discussions, and Norm Bartley for being the test subject. The authors also acknowledge the financial support provided by MICRONET, Federal Centre of Excellence on Microelectronic Devices, Circuits and Systems and the Natural Sciences and Engineering Research Council of Canada.

References

[1] H. Li, P. Roivainen and R. Forchheimer, "3-D Motion Estimation in Model-Based Facial Image Coding", *IEEE Trans. Patt. Anal. Machine Intell.*, Vol. 15, No. 6, pp. 545–555, June 1993.

[2] C. S.Choi, K. Aizawa, H. Harashima, and T. Takabe, "Analysis and Synthesis of Facial Image Sequences in Model-Based Image Coding", *IEEE Trans. on Circuits and Sys. for Video Tech.*, Vol. 4, No. 3, pp. 257–275, June 1994.

[3] M. Hötter, "Optimization and Efficiency of an Object-Oriented Analysis-Synthesis Coder", *IEEE Trans. on Circuits and Sys. for Video Tech.*, Vol. 4, No. 2, pp. 181–194, Apr. 1994.

[4] H. G. Musmann, M. Hötter and J. Ostermann, "Object-oriented Analysis-Synthesis Coding of moving images", *Signal Processing: Image Communication*, Vol. 3, No. 2, pp. 117–138, Nov. 1989.

[5] R. Koch, "Dynamic 3-D Scene Analysis through Synthesis Feedback Control", *IEEE Trans. Patt. Anal. Machine Intell.*, Vol. 15, No. 6, pp. 556–568, June 1993.

[6] A. Eleftheriadis and A. Jacquin, "Model-Assisted Coding of Video Teleconferencing Sequences at Low Bit Rates", *Proc. IEEE Intl. Symp. on Circuits and Sys.*, pp. 177-180, 1994.

[7] J. P. Lewis and F. I. Parke, "Automated Lip-Synch and Speech Synthesis for Character Animation", *Proc. CHI+GI*, pp. 143–147, 1987.

[8] D. R. Hill, A. Pearce and B. Wyvill, "Animated speech: an automated approach using speech synthesized by rules", *The Visual Computer*, Vol. 3, pp. 277-289, 1988.

[9] K. Waters, "A muscle model for animating three-dimensional facial expression", *Computer Graphics*, Vol. 21, No. 4, July 1987.

[10] D. Terzopoulos and K. Waters, "Analysis and Synthesis of Facial Image Sequences Using Physical and Anatomical Models", *IEEE Trans. Patt. Anal. Machine Intell.*, Vol. 15, pp. 569–579, June 1993.

[11] C. L. -Y. Wang and D. R. Forsey, "Langwidere: A New Facial Animation System", *in Proceedings of Computer Animation '94*, pp. 59–68, 1994.

[12] D. H. Klatt, "Review of text-to-speech conversion for English", *J. Acoust. Soc. Am.*, Vol. 82, pp. 737-793, Sept. 1987.

[13] E. F. Walther, *Lipreading*, Nelson-Hall, Chicago, IL, 1982.

[14] P. Ekman and W. V. Friesnen, *Facial Action Coding System*, Consulting Psychologists Press, Palo Alto, CA, 1978.