D.A.V.O.S: A COMPUTERIZED VIDEODISC ANIMATION WITH SYNCHRONIZED SPEECH

Paul Girard

Université du Québec à Chicoutimi,
Section Informatique, D.S.E.A.,
555 blvd Université, Chicoutimi, P.Q,
Canada, G7X 9H1

ABSTRACT

This paper presents a new approach concerning computer animation and friendly user interface. An automated method of creating a real-time videodisc animation with real or synthetic actors synchronized with a spoken text is described.

In this method, a text file is first created by a standard editor or produced by another program (e.g. the intelligent tutor); then it is translated to input a speech synthesizer and a videodisc drive in order to produce an audio-visual animation. Real-time synchronization is assured by the reception of indexes between words by a synchronizing CPU from the videodisc and the synthesizer processors. Real-time external user control of the synthesizer or the videodisc parameters is also possible during the animation via a user interface and the use of an open architecture.

RÉSUMÉ

Ce papier présente une nouvelle approche concernant l'animation par ordinateur en vue d'une interface homme-machine amicale. Cette méthode automatique permet de créer une animation en temps réel à partir d'acteurs réels ou synthétiques synchronisée avec un texte parlé.

Dans cette méthode, un texte est d'abord créé par un éditeur standard ou produit par un programme quelconque comme par exemple, un didacticiel intelligent. Il est ensuite traduit afin d'alimenter un synthétiseur de parole et un vidéodisque de façon à produire une animation en temps réel. La synchronisation en temps réel est assurée par la réception d'index spéciaux transmis par les processeurs du synthétiseur et du vidéodisque et insérés entre chaque mot du texte auditif et visuel. Un contrôle externe avant et pendant l'animation est possible grâce à une interface et un protocole approprié entre l'ordinateur de l'usager et DAVOS.

Keywords: videodisc, synthesizer, animation, synchronization, man-machine interface, real-time.

INTRODUCTION

It is an accepted principle among professional trainers that seeing and hearing produces better retention in trainees than only hearing (lectures, tapes) or only seeing (text, graphics) [NEWT85]. One way to improve the impact of a message is to "humanize" it by animating a model of a human face synchronized with the spoken words of the message; such a system should be as flexible as a standard text editor.

Several synthetic computer models of the human face have been developed over the last 15 years [LEPA87], [PWWH86], [BELA85], [PARK74], [PARK75], PARK82], [PLBA81]. Some of them can be animated but none of them could replace a real actor seen on television at 30 frames/second. The personality and human expressions almost disappear with a synthetic face; the way we communicate is often more important than the message itself.

The approach taken by Daniel Thalmann and N. Magnenat-Thalmann is probably the most realistic because their synthetic actors are produced from a digitized image of real actors manipulated by a specialized graphics software. However, this tool is more adapted for cinematography animation because the speech and the animation are not made in real time and the development cost is very high [THAL87a], [THAL87b], [THAL88].

As a result of recent technological developments, many videodiscs now can function under internal and external computer control. However, most of them are not fast enough to provide a controlled animation of 30 frames/second from a set of predetermined frames because a jump from a frame to a non-sequential one will desynchronize the

video output for a period of approximately 0.3 to 5 seconds depending on the number of tracks to be jumped and the characteristics of the drive. Videodiscs are good if the waiting time between jumps is short compared to the playing time, and this is certainly not the case if someone wants to break the normal sequential video sequence 30 times/second.

When the desired speech is specified in a textual rather than auditory form the quality of a rule-based synthetic speech is acceptable [LEPA87]. Although a small number of applications have been satisfied by the use of a barely intelligible voice, the majority of voice output applications (e.g. Computer-Assisted Instruction) requires a natural sounding voice if the man-machine interface is going to be truly user-friendly [GREE84]. Synthesized voice has improved greatly in recent years; now it has lost much of its robot-like quality [NEWT85].

This paper describes a new kind of friendly user interface which is natural and cost effective. An experimental prototype called D.A.V.O.S. has been created in order to animate a text in real time.

OPERATION OF D.A.V.O.S.

D.A.V.O.S. is an acronym for **D**istributed **A**udio-**V**isual **O**perating **S**ystem and uses four processors (see figure 1):

- an IBM PC, 256K, 2 serial ports, 1 non-standard parallel port;
- a speech synthesizer from Infovox;
- a videodisc VP935 with an internal dedicated processor;
- an external user computer running the application program such as computerized tutoring.

The data and the programs are distributed among these cooperating processors. Figure 2 shows the main functions of each one. DAVOS "humanizes" a textual message in the following manner:

1. The text is sent by the user computer at 4800 bps to a serial port of the IBM PC or it is read on a local file. A real or a synthetic actor is also selected by the user. Figure 3a shows an example of a text file.

2. The ASCII text is indexed by inserting index markers before and after each word, then sent to the speech synthesizer via a second serial port at 4800 bps. The text is translated to a phonetic form (figure 3b) and sent back to the IBM PC. The elapsed time between the reception of index markers determines the number of video frames to be allocated for each word (figure 3c).

3. The phonetic text is then translated automatically into a visual form (figure 3d) based on [GARR74] and [INJT82]. These authors identify 10 visually distinctive mouth positions and their corresponding sounds (figure 4); nine additional positions were added to assure a better video resolution by interpolating certain transitions.

4. The visual text is then converted into a indexed videodisc command file which will be used for the video animation of the selected actor (figure 3e).

5. This command file can also be edited manually to insert optionally appropriate expressions compatible with the message: laughing, surprise, eye movement, nodding, etc... The speech synthesizer will not work during these short animations (1-2 sec.); however, the two audio tracks available on the optical disc could be activated to stress these expressions. Most of these commands are "Instant Forward Jump" or "Instant Backward Jump" relative to the current video frame. Even the "Play Forward" and "Play Backward" are simulated by the use of instant jumps because each video field must be under computer control (figure 3f).

6. As soon as the speech file and the video file are produced, DAVOS initiates a real time animation where the user will see and hear the text on a video monitor, the audio part coming either from the synthesizer output or from the optical disk audio tracks. For example, in a Computer Assisted Instruction environment, a chosen professor could say to "his" student with an impatient voice after a negative evaluation: *"Jean, as-tu étudié la théorie de la relativité?"* (figure 2).

7. Different parameters regarding the synthesizer (pitch, volume, pitch variation, speech rate, degree of aspiration, type of voice, etc...) can be modified online by the user through an external computer. These commands are processed after each sentence of the animation process.

THE VIDEODISC

DAVOS uses the Philips VP935/37A parallel interface drive; it is intended for operation under real time computer control and allows for video synchronous branching of picture sequences ("Instant Jumps"). In 1986, this equipment was the only one, to the best of our knowledge, that truly supported consecutive instant

jumps without losing the video synchronization.

If 2 selected points are separated by 99 tracks or less (one video track = 1 video frame on a CAV (Constant Angular Velocity) disk), the instant jump command gives the possibility of going backward or forward in a visually undetectable (or seamless) video branching [PHIL85]. To produce a continuous animation at 30 non-sequential frames per second, a set of at most 20 contiguous frames must be accessed because this player supports a maximum of 600 jumped tracks/second. However, 19 video frames are sufficient to guarantee a good video resolution for this project.

The advantage of using this technology is the absence of limitation in the face model, which can be synthetic or real, and the fact that animation is done in real time.

CREATION OF THE OPTICAL DISK

Since the process of creating a disk is lenghty, expensive and irreversible, and since preliminary experiments can not be made, we decided to use a number of actors in different conditions. Fifteen sets of face models were filmed with a 16 mm camera; one set was created using a computer face model and a simple graphics editor and 14 were real actor faces (8 men and 6 women). This model was edited 10 times to represent the basic mouth positions (figure 4) and the 14 actors had to say a short sentence of 6 words (*"Jako, Rio, t'en fait pas, cela"*) representing the whole set of French visually distinctive mouth positions and their interpolations. Great care must be taken to prevent any undesirable head or eye movement during the 2-4 seconds necessary to say these 6 words. An unseen head support during this part should be used. People were asked to speak slowly and the camera was set at 48 pictures/second; the purpose was to have a better chance to catch the exact basic phonemes and a greater choice among the interpolation frames between these corres-ponding mouth positions.

An emotion is often conveyed by gestures and/or accompanied by a modification of some speech parameters (speech rate, volume, intonation, aspiration, etc…). To incorporate this human aspect in the animation process, we took the following approach: each one of the 14 individuals answered six questions in a non-verbal manner but natural sounds were permitted and recorded (laughing, breathing out, clapping, etc…). The purpose was to trigger a particular behavior: approval, dis-approval, nodding, surprise, anxiety, etc…; this is very important because each individual behaves differently. For example, a tutorial could send to the student a non-spoken message showing an approval. These short sequences (1-2 seconds) were identified and coded in order to be inserted manually (or automatically in the next phase) in the video command file to reproduce the personality of an actor. The position of the actor facing the camera was the same before each question in order to simulate a natural movement after or before a sequence of synchronized speech.

When these sequences are animated, they are played forward first and then backward in order to achieve a natural transition with the animated speech. Figure 2f shows an example of commands used in an expressive animation. Simultaneously, a modification of the speech parameters (volume, basic frequency, etc…) can support the "strength" of the message or the degree of its emotional impact.

Each of the 19 frames (corresponding to a certain mouth position) and the individual expression video frames were manually selected and transferred to a one inch master video tape and sent to an optical disc manufacturer. In the future, CD-WORM ("Write Once Read Many") technology will be used instead of a 16mm film; these new compact optical discs should accelerate the production of a master tape at a lower cost and will enable some experimenting before final commitment.

THE SPEECH SYNTHESIZER

A Text-To-Speech (TTS) synthesizer does not reproduce a specific voice but imitates a parameterized human voice. Some high-end TTS's produce a natural voice. Many companies support American English but the choice is very limited for the French language. Ideally, the synthesizer desired for our application would give in real time timing information on the duration of phonemes. A compromise solution in the absence of the ideal model is to synchronize the animation at the word level by the use of index markers.

DAVOS uses the only synthesizer that could suit its requirement at that time: the VoxBox made by Infovox, a Swedish company. It can support eight languages simultaneously: American English, British English, French, German, Spanish, Italian, Swedish and Norwegian. It can be controlled by an external computer using a set of commands executed in high priority. These commands control the intonation, pitch level, pitch variation, selection of a basic voice, aspiration control, language selection, speech rate, reading mode, user lexicon selection, speech output halt/restart, ASCII or phonetic text input, sentence saving/retrieving, index markers control and XON/XOFF or RTS/CTS protocol. A user or a

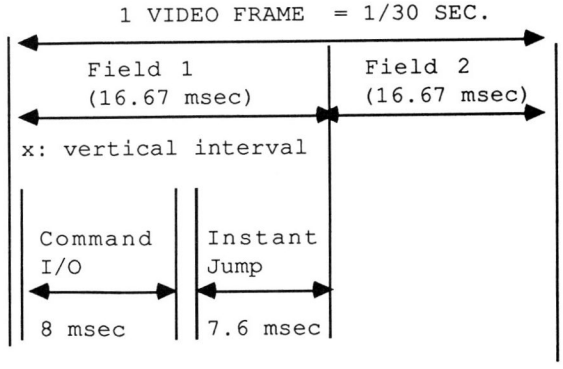
programmer can easily change any of these parameters in order to fit the desired audio message.

The VoxBox uses a MC 68000 processor and its software is stored on EPROM. This interface is well suited for a synchronized and distributed application and especially if the messages are to be said in different languages (e.g. in a computer science course, many words are said in English even if the course is in French). A few hundredths of second are sufficient to switch between two languages. The following sentence would be pronounced correctly: "*Je vais vous expliquer la différence entre un Data Base et un Data Base Management System d'après James Martin*".

SYNCHRONIZATION

DAVOS controls two asynchronous processes (the user computer and the speech synthesizer) and one synchronous process (the videodisc operation at a frequency of 30 cycles/second).

A video field (2 fields/frame) takes 16.67 msec. Approximately 8 msec are taken by the videodisc processor to transmit its own status (6 bytes), accept and validate a new video command from DAVOS (3 bytes). If the command is an instant jump command, it is started immediately and the jump will be completed before the next video vertical interval (just before video line 16 is read) in a time shorter than 7.6 msec. After the command has been accepted by the videodisc processor, DAVOS dedicates the IBM PC to the two asynchronous processes for the next 28 msec. The following schema illustrates this timing.



The speech synthesizer supports the possibility of being synchronized with another external process by the use of index markers inserted in the input text. The indexes are sent back to the host when the speech unit which follows the index begins to be spoken. It is then possible to synchronize the beginning of each spoken word with the video frame of one mouth position. The number of frames allocated to each spoken word was already measured when the phonetic text was received by the synchronizing CPU. Dectalk™ (from Digital Equipment Corporation) and CallText™ (from SpeechPlus Inc.) offer the same characteristics but they do not support French.

Four circular lists are used as temporary buffers by an interrupt service routine to assure the reception and transmission of characters from the two serial ports of the IBM PC (linked to the user computer and to the speech synthesizer). The processing of the two lists attached to the synthesizer is done during the 28 msec of free CPU time in each video frame. The user computer lists are treated after an end of sequence ("?","!" or "."). The time taken by the interrupt routine is less than 150 µsec which is fast enough to prevent a loss of the video synchronization signal (max. 600 µsec).

EVALUATION

DAVOS is an experimental prototype aiming at testing what can be achieved with available videodisc and TTS technology. Results proofed the faisability of this audio-visual animation in real time. This system provides a friendly user interface that can be used in many applications:

1. DAVOS will first be used as an output interface for the research project HERON, an intelligent tutoring system being developed at the Université de Montréal, the Université du Québec à Chicoutimi, Bishop's University and the Royal Military College of Kingston. It is still in the development phase, so the final interface with DAVOS is not completely defined yet. Actually, DAVOS is working as a stand alone system where the input message is generated by a program inside the IBM PC; a MacIntosh and an Amiga are used as the user computer to simulate the future interface.

2. DAVOS could be implemented in the majority of large companies as a tool for an audio-visual electronic mail service because a two-sided disc can hold more than 5000 sets of face models without expression or 284 sets of face models each having six sets of expression of 60 frames (2 seconds of animation). This could be done without changing the hardware already used: the frequency band of the network would not affected because the video and the audio parts are constructed locally.

3. DAVOS could also be used by a tutor oriented in the auditory training and speechreading instruction for an hearing-impaired individual. In this case, the student could optionally become his own instructor.

Actual results are very encouraging; some parts could be improved and concerns the following areas:

- the coding of non-verbal expressions;

- the classification of the speech parameters in order to facilitate the task of a user wanting to send a message of "degree x" to another user;

- the support of other languages;

- the final definition and the implementation of an external interface;

- the support of the CD-WORM (Compact disc, Write Once Read Many) technology in order to facilitate the selection of the basic mouth positions and

- the development of an auto-corrector at the field level to fix the occasional error of an instant jump command.

DAVOS is a tool that can be used in many applications limited only by the user creativity.

ACKNOWLEDGEMENTS

First, I would like to thank my advisers Gregor V. Bochmann and Jan Gecsei for their financial and intellectual support. I would also like to thank Philips Subsystems and Peripherals, Inc. for lending us the video-disc VP935/37A and especially their chief engineer Franz H. Raetzer for his technical help. I appreciated the good support given by Peter Olausson working for the Swedish company Infovox.

References

[BELA87] Bergeron,P.; Larochelle, P.; *"Controlling Facial Expressions and Body Movements in the Computer-Generated Animated short "Tony de Peltrie"*, SIGGRAPH Tutorial Notes, 1985.

[GARR74] Garric, J.; *Lecture sur les lèvres et Conservation de la Parole*, Editions Médicales et Universitaires, Paris, 1974.

[GREE84] Green, R.W.; *"Developments in Synthetic Speech"*, Sensor Review, vol 4, no 4, oct 1984, pp. 188-190.

[INJT82] Istria, M.; Nicolas-Jeantouse, C.; Tan-boise, J.; *Manuel de Lecture Labiale*, Masson 1982.

[LEPA87] Lewis, J.P.; Parke, F.I.; *"Automated Lip-Synch and Speech Synthesis for Character Animation"*, CHI + GI 1987, pp 143-147.

[NEWT85] The New Tech Training Newsletter, oct. 1985, issue #3, John Victor (Ed), TALMIS (Publ.), New-York.

[PARK74] Parke, F.; *"A Parametric Model for Human Faces"*, Ph.D dissertation, U. of Utah, 1974.

[PARK75] PArke, F.; *"A Model for Human Faces that Allows Speech Synchronized Animation"*, Computer & Graphics, vol. 1, 1975, pp. 3-4.

[PARK82] Parke, F.; *"Parameterized models for facial animation"*, IEEE Computer Graphics & Applications, 9, 2, nov 1982, pp. 61-68.

[PHIL85] *VP935 Professional Videodisc Drive Product Specification*, Philips Subsystems and Peripherals, Knoxville, Tennessee, 1985.

[PLBA81] Platt, S.M.; Badler, N.I.; *"Animating Facial Expressions"*, ACM Computer Graphics, SIGGRAPH 1981, 15(3), pp. 245-252.

[PWWH86] Pearce, Andrew; Wyville, Brian; Wyvill, Geoff; Hill, David; *"Speech and Expression: A Computer Solution to Face Animation"*, Graphics Interface '86: Vancouver, B.C., Publ. Toronto, Ont., Convention Information Proceeding Society, 1986.

[THAL87a] Magnenat-Thalmann, N.; Thalmann, Daniel; *"The Direction of Synthetic Actors in the Film Rendez-Vous à Montréal"*, IEEE Computer Graphics & Applications, vol 7, no. 6, dec. 1987.

[THAL87b] Magnenat-Thalmann, N.; Thalmann, Daniel; *"Abstract Muscle Action Procedures for Human Face Animation"*, The Visual Computer, vol. 3, no. 6, dec. 1987.

[THAL88] Magnenat-Thalmann, N.; Thalmann, Daniel; *Synthetic Actors: Computer Animation in Cinema*, Springer-Verlag, Heidelberg (to be published in 1988)

APPENDICE

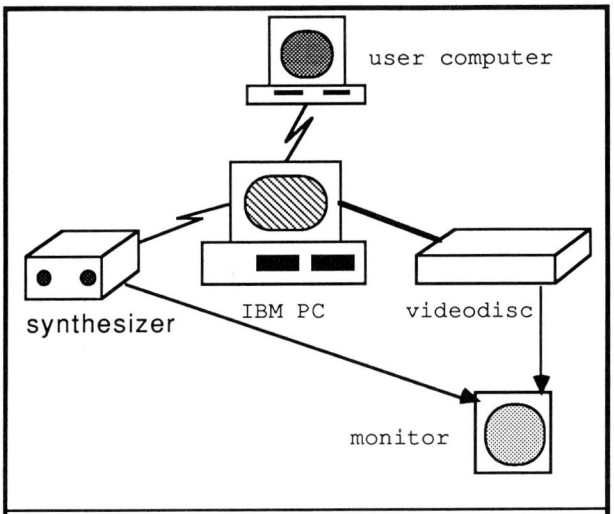


Figure 1
Physical Schema

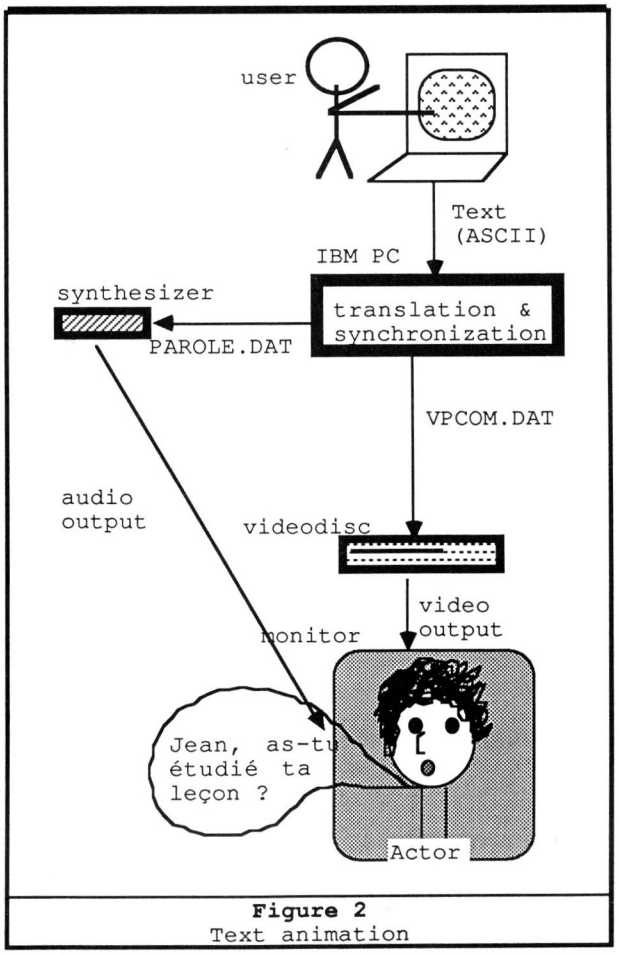


Figure 2
Text animation

```
JE SUIS UN ORDINATEUR QUI PARLE.

EXCUSEZ-MOI, J'AI UN ACCENT FRANÇAIS.

COMMENT T'APPELLES-TU MON AMI?
```

Figure 3a
Example of an input text file

```
ZH'E0+ SJ1'I+S '\9+N+ ORDINAT'\R K'Ip1+
P'ARL .

E1KSKYZ'E+Z MW'A , ZH'I ap 'E+ '\9+N+
AKS'A9+T FRA9S'E1+S .

KOM'A9+T T'E ap AP'E1L+S T'Y+ M'O9+N+
AM'I ?
```

Figure 3b
Phonetic input text
The first word "JE" has 2 phonemes: **J** and **E**.
J is represented by **ZH** and **E** by **E0**. The symbol ' represent a longer phoneme duration on the **E0** and the "+" can affect the sounding of the next word.

```
^A001JE^A002SUIS^A003UN^A004ORDINATEUR
^A005QUI^A006PARLE^A007.

^A008EXCUSEZ^A009MOI,^A010J'AI^A011UN
^A012ACCENT^A013FRANÇAIS^A014.

^A015COMMENT^A016T'APPELLES^A017TU^A018
MON^A019AMI^A020?
```

Figure 3b
Indexed input file sent to the speech synthesizer (**^Axxx**: index number). Index markers are inserted before and after each word. For example, ^A001 will be sent back to the IBM PC by the VoxBox just before the pronounciation of **ZH**.

^**A001**0702^**A002**090304^**A003**0903^**A004**0802080
408010803^**A005**04^**A006**050110^**A007**.

^**A008**0109030904^**A009**050301^**A010**0704^**A011**
03^**A012**08010902^**A013**06020901^**A014**.

^**A015**020502^**A016**0801050110^**A017**0803^**A018**0
503^**A019**08010504^**A020**.

Figure 3d
Indexed visual mouth positions file. The word **JE** or the phonetic word "**ZH E0**"is translated in the visual form "**07 02**"; in this particular case, each phoneme is visually distinctive. The class **07** corresponds to the "**CH**" or "**J**" phonemes and the class **02**, to the **AN**, **AH** and **E** phonemes.

^BD01211.
^**A001**00E00600E00600F00500F00400E00000E000
^**A002**00E00300E00000F00400E00000E01400F110
0E00000E000^**A003**00E00100E00000E00000F0040
0E00000E00000E000^**A004**00E00600E00000E0010
00F01500E01500F01000F00300E00000E00300E00
000E00600F01300E01400F00700F00600E00000E0
0000E00000E00000E00000E000^**A005**00E01400E0
0000F01100E00000E00000E000^**A006**00E00800E0
0300F01200E00000E01300F00400F01000F00200E
00900E00000E00000E00000E00000E00700E00000
F01700E000^**A007**.

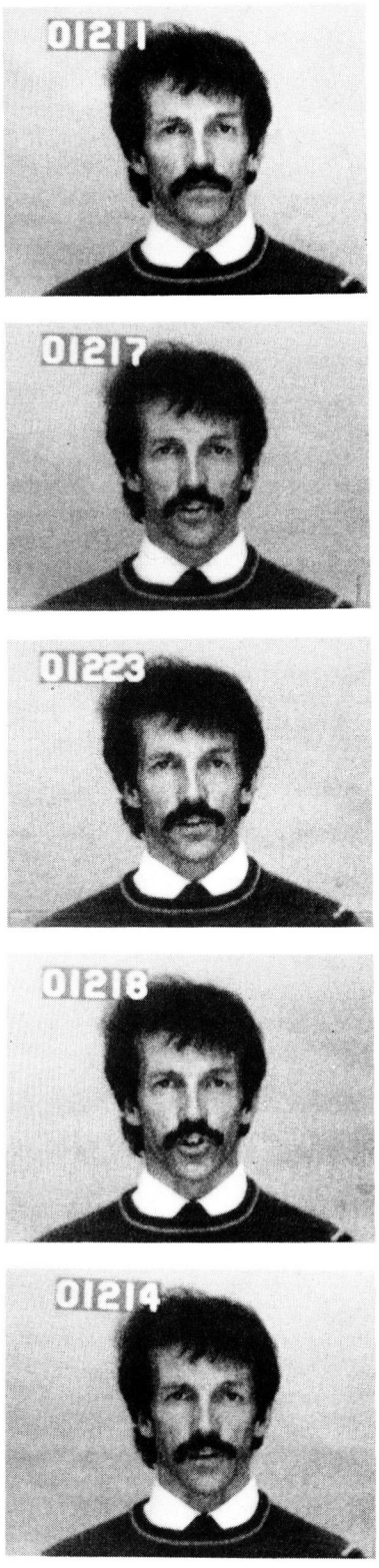
Figure 3e
First sentence of the indexed videodisc command file. The mouth positions **07** and **02** are translated to videodisc commands. Interpolation frames are added to assure a better video resolution. The command **00E006** is for an instant forward jump of 6 frames and the command **00F005** is an instant backward jump of 5 frames. One command is executed every two video fields in order to visualize each mouth position for a continuous animation.

Class	Group	French phonemes
1	a	a, è=in
2	an	an, ah, e
3	o	o=on, ou=u, eu
4	i	i, é
5	p	p=b=m
6	f	f=v
7	ch	ch=j
8	t	t=d=n, gn, ille
9	s	s=z
10	L	L

Figure 4
Classes of visually distinctive mouth positions corresponding to French phonemes.

Code	Meaning
D01211	Go to the beginning of mouth position
00E070 00E045	115 tracks are jumped
020602	Activate audio track 1
^B+30	Play forward 30 frames
^B-30	Play backward 30 frames
020600	Deactivate audio track 1
00F045 00F070.	Go to the beginning of mouth position and continue the speech synchronized animation.
etc…	
	Normal instant jumps command follow

Figure 3f
Indexed Command File with expressive sequences. In this example, the laughing sequence is 30 frames long and located 115 frames after the beginning of the basic mouth position frames.



Visual animation of the word JE.
The optical disk address is shown
on the video frame.