A System for Multimodal Speech Reproduction

NICOLAU L. WERNECK[1], LUCAS R. P. MALTA[1], LEONARDO C. ARAUJO[1] AND HANI C. YEHIA (ADVISOR)[1]

[1]CEFALA — Center for Research on Speech, Acoustics, Language and Music
{nwerneck,malta,leoca,hani}@cefala.org
UFMG — Universidade Federal de Minas Gerais, Belo Horizonte, MG, Brazil

Abstract. It is described a program created for the production of realistic animations of tridimensional facial models in real time, based on LPC analysis of speech and previous determination of the characteristic dynamics of the speaker model. The final program was successfully able to render a model with 900 triangles in a frame with 350x400 pixels at 60 frames per second, driven only by parameters extracted from a speech signal.

1 Introduction

Speech is usually understood as an acoustic process, but it has been proved that listeners also acquire visual information during a dialogue [1][2]. Speech perception is a bimodal process, in which both auditory and visual perception play their roles. A striking demonstration of this fact was discovered when Harry McGurk and John MacDonald were studying how infants perceive speech during different stages of development and accidentally created a videotape with the audio syllable /ba/ dubbed onto a visual /ga/. When listeners watched the tape they perceived /da/, which is articulatorily between these two. This audio-visual illusion has become known as the McGurk effect [3][4].

The efficiency of speech-based communication can be improved by showing the speaker's image together with the voice signal [1]. This means that there are real benefits on the audio-visual transmission of speech, and not just pointless science fiction. Speech communication systems that show the image of a speaker together with the sound of his voice are being called *multimodal* systems by researchers.

Our system is focused on the generation of the image of a speaker driven by the acoustic information transmitted, in order to enhance the realism of a spoken signal. The system also finds application on general computer animation of characters, although this activity does not take advantage of the high speed attainable by it.

The basis for the system was described by Yehia, who demonstrated that it is possible to estimate facial movements from speech signals [5].

The most important characteristic of the system is the use of the Line Spectrum Pairs (LSP) analysis, which is largely used in speech research, and turned up to be very fit to be used as a source to the face modeling program. It is convenient to use this analysis because it is directly related to the Linear Predictive Coding (LPC) parameters that can be used to synthesize speech as shown below.

2 Methods

2.1 General Process Description

There are two different data inputs to the system: the speech signal and a previously determined face model that contains information about the shape of the speaker and how it moves. The speech signal is analyzed in frames, where there are extracted a set of parameters to be transmitted and used to resynthesize each speech segment, as well as to control a program that renders the reshaped face to be displayed with the current acoustic signal (Figure 1).

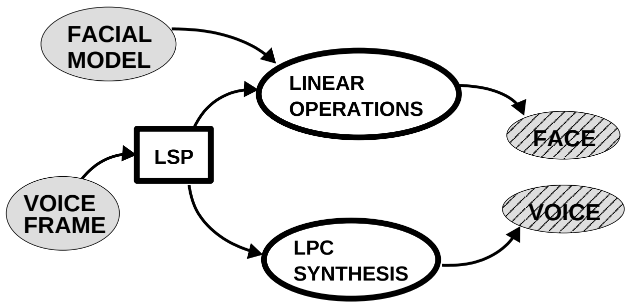

Figure 1: Process diagram.

In the figure 1 it is depicted a facial model used by the system and a voice frame that is taken from a speech signal. The voice frames are analised, and parameters are extracted, which are then transmitted to the face and speech synthesis programs. The picture shows a block with a LSP analysis, explained further in this text. In the system described here, this analysis yields 12 floating point numbers from a speech frame of 134 samples. The parameters extracted then feed the system, yielding a single animation frame and audio frame from each recorded speech frame. The audio signal is created by LPC synthesis, also explained further here. The linear opeartions are deformations done in the facial model based on linear relations, and could be substituted by more elaborate non-linear functions.

This process has already been implemented [6][7][8], but that system could only work from previously recorded signals because the process was very slow. In this work the process is recreated with a stand-alone program that works faster, animating the model in real time.

2.2 Model of the Face

The facial model that must be previously created to be used by the system contains the following information:

- Mesh and applied texture.
- Mean of the input parameters.
- Linear weights matrix.

The process of acquiring this information is done on two steps. First the head of the speaker is digitized on a 3D-scanner on different facial configurations. These configurations must explore all the movements that can happen independently during normal speech. If a small number of configurations are used, the system will not be able to do a realistic simulation, and if too much are selected, than the model will contain more degrees of freedom than it is needed to simulate only normal speech.

After the head is acquired at these different states, the resulting models are still unrelated objects. The meshes must be transformed so that each vertex in a model corresponds to vertices in each of the other models, and so any mesh can be recreated from any other by simply moving its vertices around. Each mesh becomes then a distorted version of the others. This way it is possible to represent the different configurations of the face as vectors in a space where the dimensions are the coordinates of the points of the meshes. The final mesh that is stored in the model is the mean of the acquired faces.

In the second step, some of these vertices are identified as being control points previously marked on the face. To create the model used in this work it was used 18 control points. A linear model is then computed to determine the displacement of the vertices of the mean facial model $\vec{\mathbf{M}}$ from the displacement of the control points to their mean state $\vec{\mathbf{C}}$. This is the linear weights matrix $\mathbf{W}$ that can be used to create a new face $\vec{\mathbf{M}}_{cur}$ from a vector of current control points positions $\vec{C}_{cur}$ in the following way:

$$\vec{\mathbf{M}}_{cur} = \vec{\mathbf{M}} + \mathbf{W}(\vec{\mathbf{C}}_{cur} - \vec{\mathbf{C}}) \qquad (1)$$

The same control points are then monitored with a 3D motion capture system while the subject pronounces various sentences. This motion of the points is then analyzed together with the recorded audio signal of the sentences, and a model is created relating acoustics and facial motion during speech production. This model can be used to estimate the positions of the control points from different parameters extracted from speech signals. The mean of these parameters is the last piece of the speaker model.

This step restricts the model to be used only in speech applications, as the face is not monitored in all of its possible states, but only in the subset of states that are related to speech. In other words, this model cannot be used to create an animation of a face doing different expressions, or doing movements that do not occur during normal speech. For example, all movements tend to be symmetric during speech, but this is not a restriction of human faces in general. While these restrictions make the model useless for general animations, it also makes it more simple.

Using functions to determine the control point movements from speech parameters, and the rest of the face from the control points, it is possible to determine the whole facial and head motion from the analysis of speech.

The mapping from the control to the facial points was successfully done by linear relations, but the mapping from the speech parameters to the control points show better results with non-linear functions [5]. This work was done with a completely linear model, but an extension to apply neural networks to the system, as done by Yehia [5], has been developed and is under tests. Any non-linear estimator can be appended to the input of the face rendering program, but this program itself uses only a linear model.

It has not yet been implemented a program to manage the creation of the speaker model. The model used here was computed by an *ad hoc* process.

This kind of multimodal speech reproduction system is not limited to 3D models. Barbosa and Yehia have already demonstrated the possibility to restrict the process to two dimensions [8]. The system this article describes is capable of handling such a model, but does not take any advantage on the dimensional restriction imposed.

2.3 Sound Processing

The speech signal is divided into frames of approximately 20ms, where it may be considered stationary [9]. Each frame is analyzed in order to extract the parameters that will be transmitted and used on the receptor to synthesize the sound and the face. These parameters are the root mean square of the signal and the LSP (Lines Spectrum Pairs) parameters computed from a tenth order LPC (Linear Prediction Coding) analysis [10][11][12]. These analysis are defined below.

2.3.1 LPC analysis

Vocoders are a class of speech coding systems that analyze the signal at the source and transmits to the receiver a set of parameters that control a voice synthesizer. Because vocoders take advantage on physical restriction of the modeled signals, they use to be more complex than waveform coders and archieve very high compression rates.

A popular vocoder technique is the LPC, developed by Itakura [11] and Atal [13] independently. In a LPC system, speech is the output of a digital filter having as input either an impulse train or noise, depending on whether the speech segment is voiced or not. This procedure emulates the production of human voice. For each frame the only values needed to be transmitted are the LPC coefficients, the gain and the pich. Pitch zero means an unvoiced frame (Figure 2).

The LPC filter of the p-th order is an all-poles filter defined y the transfer function:

$$H(z) = \frac{G}{1 + \sum_{k=1}^{p} a_k z^{-k}}. \tag{2}$$

The coefficients of this filter can be determined using any linear autoregressive modeling technique.

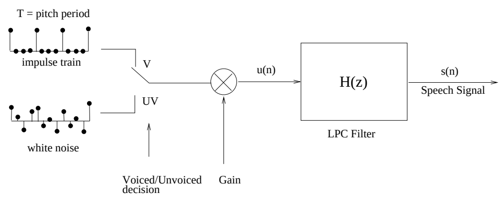

Figure 2: LPC synthesizer.

The LSP parameters further developed by Itakura [10] are a mapping of the LPC parameters, usually computed by finding the roots of two polynomials defined from the LPC values. LSP values bear a close relationship with the formant frequencies of the vocal tract. They are more stable in time and have other characteristics that make them more suitable than LPC to various applications. This was the case in the problem of the determination of facial behavior from speech signal [5]. The mapping between LPC or PARCOR [10] and facial motion is much more complicated than the mapping with LSP parameters.

The system has been programmed to use the autocorrelation method with the Levinson-Durbin algorithm to calculate the LPC coefficients. To calculate the LSP from the LPC values it was implemented the algorithm created by Huang Dezhi[12].

2.4 Computer Graphics Modeling

The development of graphics libraries has been an important factor in spreading the use of computer graphics. Those libraries provide an underlying portable software platform that optimizes the utilization of the available graphics hardware. In this context, `OpenGL` has become a standard API for intensive graphics applications. In this work it suited our needs because of its high performance on appropriate hardware and accessibility.

The core program of the system was done in the `C++` programming language. A library of classes to deal with the models was created. Points were stored in a vector, and the triangles list of the mesh was stored as a vector of pointers to the points. Thus the coordinates of the vertices were calculated only once for each frame. Each vertex also stored the linear weights associated to it, and also the direction of a normal to do lighting with Gouraud shading [14]. To read the model file into those classes, a parser program was created with the `lex` language (Figure 3).

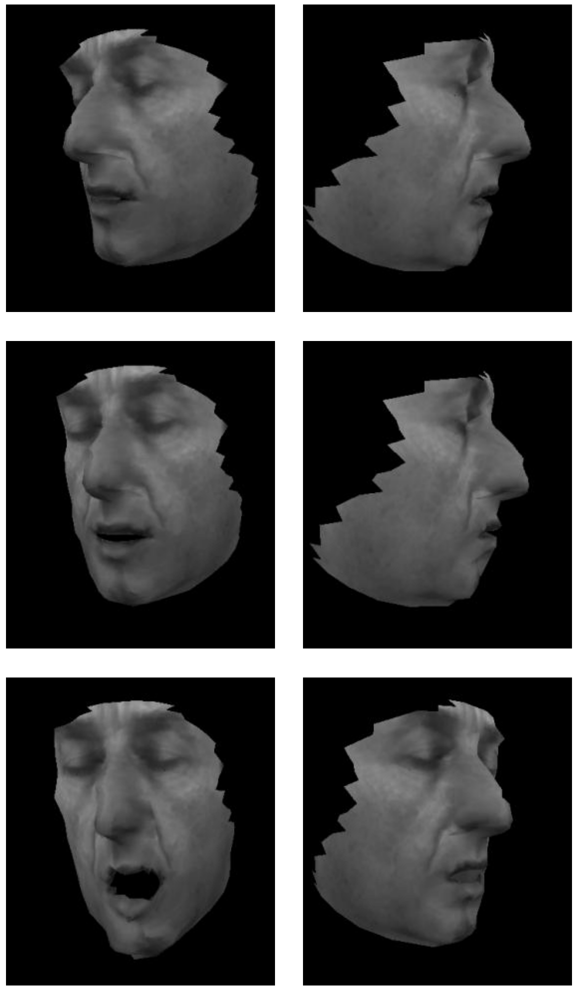

Figure 3: Example faces. These frames were selected from an animation created by the system based on a real speech signal, with the model facing different directions.

When the program begins, the model is read to the memory, and a window is opened with the model properly placed on the center of the image. Using command keys, the user can rotate the model clockwise or counterclockwise in relation to the vertical axis, and can also move a light source to anywhere in the space. The current program supports only one texture, which can be a picture taken at the time the model was acquired, as was used on the example pictures shown here, or any other picture. In the future it will be used texture animation to enhance the realism of the model, allowing the blinking of the eyes and perhaps the appearence of wrinkles due to facial movement.

The program also receive commands from the standard input, which is continuously read by a concurrent thread. A command to change the position of the camera, and various other more complex commands will be included in the future to control other aspects of the scene, as lighting for example.

There is also a command that states the new parameters to be used to render the face. When this command is issued, the parameters are appended to a buffer that is repeatedly read by a function that applies the parameters to the vertices of the model. The parameters entered are the pure LSP values, and they have to be first subtracted from their mean values which are also part of the speaker model. These differences are then simply multiplied by the weight matrix to determine the displacement to be added to the coordinates of the facial model's mesh to determine the current mesh.

3 Conclusion

The final system could successfully generate animations of a model with 900 triangles at 60 frames per second in a window with 350x400 pixels using a personal computer with a NVidia 3D graphics card. This rate is the same at which speech frames were transmitted, and slower rates can be attained with appropriate interpolation of parameters from sets of speech frames.

With the proper texture, lighting and perspective, the rendered model looked very realistic.

Acknowledgements

This work was done in collaboration with ATR [1] (Japan) and partial support was provided by CNPq [2] (Brazil).

References

[1] Q. Summerfield, "Use of visual information for phonetic perception," *Phonetica*, no. 36, pp. 314–331, 1979.

[2] E. Vatikiotis-Bateson, I.-M. Eigsti, S. Yano, and K. Munhall, "Eye movement of perceivers during audiovisual speech perception," *Perception & Psychophysics*, 1998.

[3] H. McGurk and J. MacDonald, "Hearing lips and seeing voices," *Nature*, no. 264, pp. 746–748, 1976.

[4] J. MacDonald and H. McGurk, "Visual influences on speech perception processes," *Perception and Psychophysics*, no. 24, pp. 253–257, 1978.

[5] H. Yehia, P. Rubin, and E. Vatikiotis-Bateson, "Quantitative association of vocal-tract and facial behavior," *Speech Communication*, vol. 26, no. 1–2, pp. 23–43, 1998.

[6] E. Vatikiotis-Bateson and H. C. Yehia, "Speaking mode variability in multimodal speech production," *IEEE Transactions on Neural Networks*, vol. 13, pp. 894–899, July 2002.

[7] H. C. Yehia, T. Kuratate, and E. Vatikiotis-Bateson, "Linking facial animation, head motion and speech acoustics," *Journal of Phonetics*, vol. 30, pp. 555–568, July 2002.

[8] A. V. Barbosa and H. C. Yehia, "Measuring the relation between speech acoustics and 2D facial motion," in *Proc. ICASSP'2001, International Conference on Acoustics, Speech and Signal Processing*, vol. I, pp. 181–184, 2001.

[9] L. Rabiner and R. Schafer, *Digital Processing of Speech Signals*. Prentice Hall, 1978.

[10] N. Sugamura and F. Itakura, "Speech data compression by lsp speech analysis-synthesis technique," *Trans. Electronics and Communications in Japan*, no. 64A, pp. 599–606, 1981. (SSSHP 28 reprints).

[11] F. Itakura and S. Saito, "Analysis synthesis telephony based on the maximum likelyhood method," in *Speech Synthesis* (J. Flanagan and R. Rabiner, eds.), pp. 289–292, Dowden, Hutchinson & and Ross, 1973. (Rep. from 6th Int. Cong. Acoust., Tokyo, 1968.).

[12] H. Dezhi and L. CAI, "A new approach for computing lsp parameters from 10th-order lpc coefficients," *ICSP2002*, vol. 1, pp. 434–436, 2002.

[13] B. S. Atal and S. L. Hanauer, "Speech analysis and synthesis by linear prediction of the speech wave," *Journal of the Acoustic Society of America*, vol. 50, pp. 637–655, 1971.

[14] H. Gouraud, "Continuous shading of curved surfaces," *IEEE Transactions on Computers*, vol. C-20, pp. 623–628, June 1971.

[1] Advanced Telecommunications Research Institute International

[2] Conselho Nacional de Desenvolvimento Cientfico e Tecnolgico