Artificial Visual Speech Synchronized with a Speech Synthesis System

H.H. Bothe und E.A. Wieden

Department of Electronics, Technical University Berlin
Einsteinufer 17, D-10587 Berlin, Germany

Abstract: This paper describes a new approach of modeling visual speech, based on an artificial neural network (ANN). The network architecture makes possible a fusion of linguistic expert knowledge into the ANN. Goal is the development of a computer animation program as a training aid for learning lip-reading. The current PC version allows a synchronization of the animation program with a special stand-alone speech synthesis computer via a Centronics parallel interface.

1 Introduction

From the experimental work of Menzerath, together with de Lacerda [1] it is known that the movements of the speech organs are structurally interrelated within the spoken context. The speech organs needed for the formation of upcoming phones, even though currently not engaged, take up position relatively early to their actual use. They produce sound in the course of a fully overlapping phonal coarticulation. The projection of these movements on the speakers face may be seen as *visual speech*. They mostly contain sufficient information to enable hearing-impaired persons to lip-read a spoken text. The visual recognition is largely focussed on the speakers mouth region, especially on the lips. Since lip movements contain most of the visually perceptible information, this paper proposes the modeling of face movements with the help of related lip shapes only.

A realistic appraisal of the research effort leads to necessary limitations due to the high number of influencing factors (e.g. speech specific facial physiognomy, dialect, speed of delivery, sentence and word stress). On one hand, the word material on which the movement analysis is based has to be fixed on a representative subset of the existing phonetic sequences. Thus, the developed motion model is an extension of this subset. On the other hand, the investigation is limited to prototypic speakers.

The smallest speaker-independent units derived from the acoustic signal being semantically distinguishable are the phonemes. The smallest perceptible units of the visual articulatory movements are called visemes. The German language consists of ca. 40 phonemes and 12 visemes [2].

In order to model visual speech movements, the acoustic speech signal and movement data of prototype speakers were recorded on videotape and analyzed with a workstation. For automatic visual feature extraction in the speaker's face several points on nose and forehead, as well as the lip contours, were marked with a contrasting color. One example frame and the feature extraction is shown in figure 1. The two marked reference points and the set point on the nose refer to the head coordinate system.

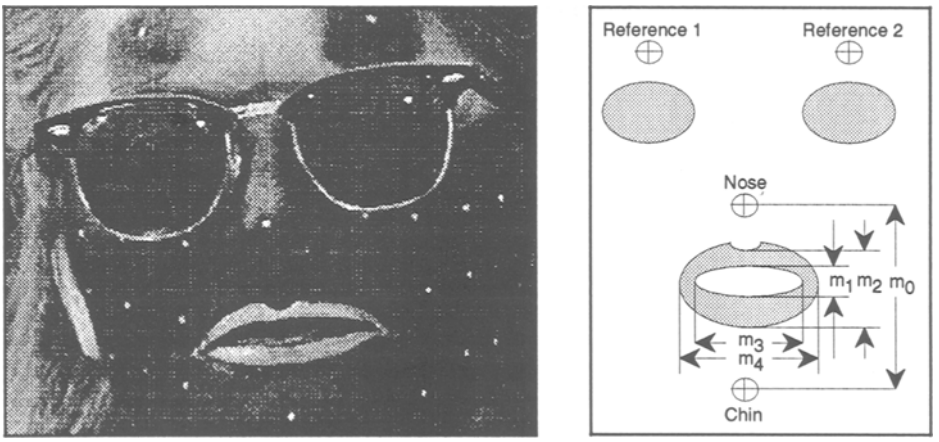

Fig.1. Typical video frame and visual feature extraction.

Those frames fitting best with the subjective impressions for a well pronounced sound were interactively indicated with the help of both the acoustic and visual material by different experts in lipreading; the acoustic phone boundaries - determined with the help of oscillogram, sonagram and playback - limit the scanning range of each wanted frame [3]. In certain cases as, for instance, for the phonemes /h,g,k/, no characteristic frame could be determined.

The proposed set points and contours were localized with the help of an automatic contrast search program. The determined characteristic frames of the text corpus were classified with respect to lip shape and position. The cluster centers compose a set of representative key-frames. In the later computer animation, the given phoneme input sequence is first mapped on a corresponding sequence of key-frames; then, the film is being generated by calculating interim frames within this framework. The linguistic data are at the same time used to control a synchronized acoustic speech synthesis computer. A block diagramm of the analysis-synthesis-system is shown in figure 2.

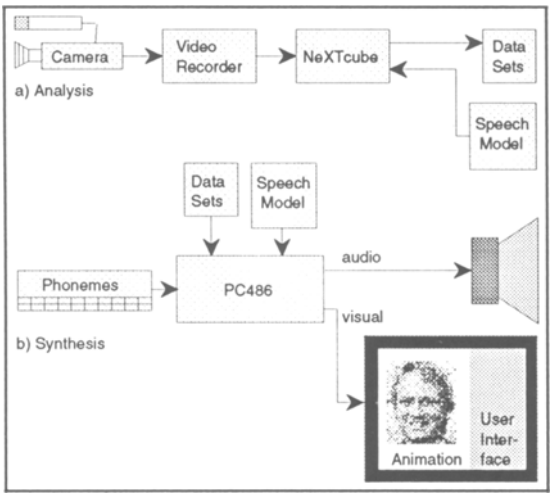

Fig. 2. Analysis-synthesis-system.

2 Key-frame Selection by an Artificial Neural Network

The subject of modeling visual speech and coarticulation effects has been addressed by several authors [4-8] for different languages. The movements are either related to an interpolation between a fixed set of key-frames or controlled by certain visual features. Key-frames may be related to the visemes of the corresponding text. A first order approximation for modeling backward and forward coarticulation effects takes into account the immediate next neighboring phonemes. For this purpose, the phonematic text is split into overlapping diphones and diphthongs are represented by two closely connected single phones, whereas the frames of the second are classified with respect to the first one [7]. This process leads to a deterministic diphone related phoneme-key-frame mapping.

In reality, coarticulation effects extend often far beyond the immediate next neighboring phonemes. A proposed area of influence has strong limits by the need of a finite text corpus.

After establishing a diphone based model, large text corpora with 84 sentences and frequently used German words have been analyzed in order to improve the quality of the motion model. Besides the integration of syllable core-zones or the phoneme representation by a variable amount of key-frames, as proposed in [8, 9], a further step to improve the model is to allow a context depending fuzzy phoneme-to-viseme mapping.

In this case, the feature vectors were classified by means of a fuzzy c-means algorithm [10]. The algorithm generates optimum location of the clusters automatically with respect to a given number of clusters. The representatives of the clusters are again taken as key-frames. In this case, there is no a priori correlation between phonemes and key-frames. For the later visual speech synthesis an ANN has been trained to select the frames with respect to the given phoneme sequence.

The developed motion model consists of a multi-layer neural network; it selects the specific key-frames related to the single phonemes and the surrounding next 3+3 neighbor phonemes. In the later speech synthesis, the film is again generated by calculating interim frames. The general design of the ANN is shown in figure 3.

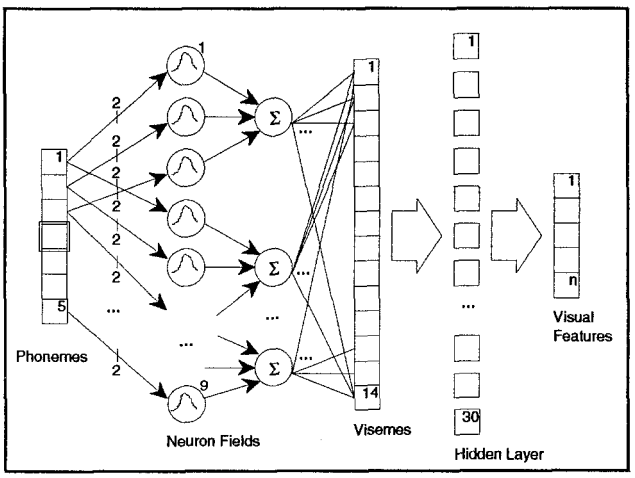

Fig. 3. Design of the ANN for key-frame selection.

In a first step the input sequence of 7 phonemes each out of the set of 41 phonemes is mapped on a set of 14 output neurons by a RBFN with Gaussian distribution functions. Each output neuron represents one viseme. This means that a certain phoneme in its context of 3+3 neighboring phonemes is represented by 14 membership grades to the set of visemes. Linguistic knowledge on the dynamics of the articulation process can be fused in before learning the RBFN, e.g. with methods described in [11].

The viseme neurons are taken as input neurons for a subsequent multi-layer perceptron (MLP) with 30 neurons in one hidden layer. The 5D output vector is pointing to the proposed corresponding feature values. The actual key-frame is selected by using the nearest neighbor method and the Euclidian distance measure.

The network is trained in three steps: i) the phoneme-to-viseme mapping with respect to the visematic system (e.g., since /p,b,m/ belong to the same viseme, a crisp mapping on the /p,b,m/-viseme neuron is proposed when /p,b,m/ are in the center position of the input sequence), ii) the viseme-to-feature-vector mapping with respect to the corresponding training sets, iii) the connected ANN with respect to the given phoneme-to-feature-vector mapping.

The ANN approach allows to i) forecast the course of features for any given input text and ii) refine the so far in the literature crisp phoneme-to-viseme mapping by taking contextual influences into account.

3 Audio-visual speech synthesis

In order to improve the ability of lipreading for hearing-impaired persons, a computer animation program may serve as a language lab for visual speech as described in [7].

In order to improve the diphone based key-frame selection, a motion model that is based on key-frame selection by the described neuro-fuzzy method has again been implemented on a PC, this time with an open amount of key-frames per phoneme. The resulting system is still in the state of pre-testing and experimentation.

The interim frames between the key-frames are calculated by a morphing algorithm. For this purpose, a fixed amount of set points have been interactively placed in each key-frame, serving as a framework for the triangulation algorithm. These set points are related to specific physiological points in the speakers face. For a given phoneme sequence, the interim frames are placed in a grid of equidistant time intervals as indicated in figure 4.

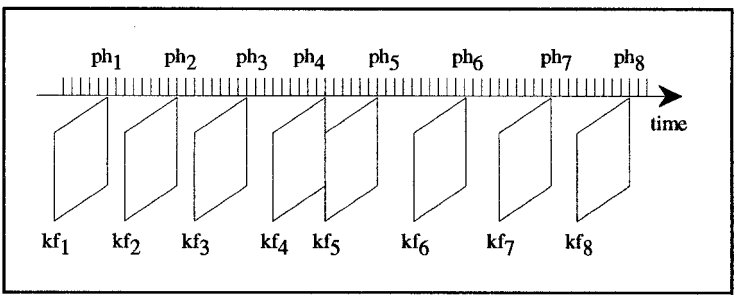

Fig. 4. Selection of key-frames *kf* and calculation of interim frames.

As an exemplary result, figure 5 shows the predicted and the original course of the visual feature m_2 = *height of the outer mouth contour*, referred to an arbitrary maximum value.

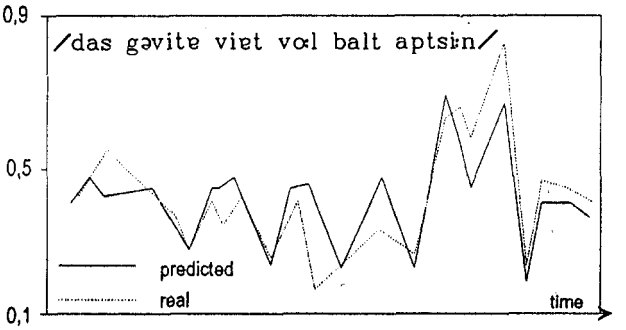

Fig. 5. Predicted and original course of m_2 = *outer height.*

The playback speed of the animation system can be varied in several steps from slow motion to time lapse. The input phoneme sequences are entered by keyboard or mouse and can be combined according to content to form corresponding lessons. The sequence to be depicted from the contents of the chosen lesson can also be selected by a random sequence generator.

The animation system has been synchronized with a stand-alone speech synthesis system which is based on a M68008 microprocessor and some specific hardware devices. For a communication that allows also slow-motion and high-speed speech, the PC as the host computer controls the communication with the speech synthesis computer via a Centronics interface. A block diagram of the speech synthesis computer is shown in figure 6. The speech synthesis uses several diphone based look-up tables.

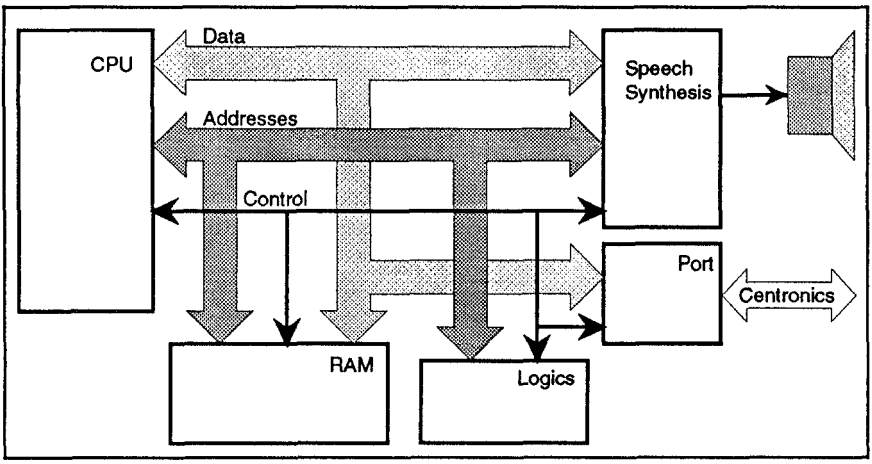

Fig.6. Block diagram of the speech synthesis computer.

Simultaneously with the animated cartoon, a time series of finger signs (dactyls) - correlated with the spoken text - can be presented in an extra window on the computer screen. The speed of the film creates the illusion of moving fingers.

References

1. P. Menzerath and A. de Lacerda: Koartikulation, Steuerung und Lautabgrenzung, Berlin, 1933.

2. G. Alich: Zur Erkennbarkeit von Sprachgestalten beim Ablesen vom Munde (Dissertation), Bonn, 1961.

3. H.H. Bothe and F. Rieger: Lipreading - Analysis and Synthesis on Microcomputers, in: W. Zagler (Ed.), Computers for Handicapped Persons, Proceedings of the 3rd International Conference, Vienna, (1992), 59-64.

4. D. Storey and M. Roberts: Reading the Speech of Digital Lips: Motives an Methods for Audio-visual Speech Synthesis, Visible Language 22 (1989), 112-127.

5. M.M. Cohen and D.W. Massaro: Synthesis of Visible Speech, Behaviour Research Methods, Instruments & Computers, (1990), 260-263.

6. M. Saintourens, M.H. Tramus, H. Huitric, and M. Nahas: Creation of a Synthetic Face Speaking in Real Time with a Synthetic Voice, Proceedings of the Workshop of Speech Synthesis, Autrance, (1990), 381-393.

7. H.H. Bothe, G. Lindner and F. Rieger: The Development of a Computer Animation Program for the Teaching of Lipreading, In: E. Ballabio, I. Placencia-Porrero and R. Puig de la Bellacasa (Eds.), Technology and Informatics 9, Rehabilitation Technology: Strategies for the European Union, Amsterdam, (1993), 45-49.

8. H.H. Bothe, F. Rieger and R. Tackmann: Visual Coarticulation Effects in Syllable Environment, Proceedings of the EUROSPEECH, Berlin, (1993), 1741-1744.

9. H.H. Bothe, G. Lindner, R. Pramanik and F. Rieger: Dynamic Modeling of Visual Articulation Movements, Proceedings of the International Symposium on Nonlinear Theory and its Applications (NOLTA), Hawaii, (1993), 1363-1366.

10. J.C. Bezdek: Pattern Recognition with Objective Function Algorithms, London, 1981.

11. J.S. Roger Jang and C.T. Sun: Functional Equivalence Between Radial Basis Function Networks and Fuzzy Inference Systems, Trans. Neural Networks, Vol. 4, No. 1 (1993), 156-159.