ARTIFICIAL VISUAL SPEECH (AVS) CONTROLLED BY FUZZY METHODS

H-H. Bothe
Technical University of Berlin, Electronics Institute, Berlin, Germany

ABSTRACT

This paper describes a new approach to modelling visual speech movements with the help of a complex fuzzy-neural network (FNN), putting a particular emphasis on the input coding. The FNN uses a set of characteristic key-pictures derived from video films with prototypic speakers. A later facial animation may be created by arranging a sequence of key-pictures with respect to the given phonetic input text and a subsequent calculation of interim frames. For this selection, The FNN consists of a radial basis function network, a multilayer perceptron and a self-organizing Kohonen map.

Goal of the described work is the development of a facial animation program for the teaching of lip-reading that may be applied in schools or rehabilitation centers for hearing-impaired people. The present version creates grey-scale films on the computer screen that correspond to a given input text. It is implemeted on PC (MS-DOS) and is prepared for additional connection to a synchronized speech synthesis computer.

INTRODUCTION

Since the experimental work of Menzerath and de Lacerda [1] it is known that the movements of the speech organs are structurally interrelated within a spoken context. The sound signals are created in the course of a fully overlapping *coarticulation*.

The resulting facial movements can be treated as visual speech. While the smallest speaker-independent perceptual units of the acoustical signal are the *phonemes*, the correlating visual speech units are called *visemes* [2].

In spite of a large data reduction visual speech contains usually sufficiant information to enable hearing-impaired people to lip-read. The largest part of information is derived from the movements of the mouth region.

Thus, this paper postulates the modelling of facial movements that are relevant for the process of lip-reading with the help of changes in the lip contours. Although the development of a general motion model on sophisticated animation computers is desirable, this work concentrates on the implementation of one realistic prototype model that can be used in schools or rehabilitation centers. The proposed motion model has to take the resulting limitations in account. Other approaches are, for instance, described in [3-5].

DATA ACQUISITION

A block diagram of the complete analysis-synthesis-system is shown in Figure 1. At first, the acoustic speech signal and video data of prototype speakers were recorded on videotape and analyzed for a text corpus of 84 sentences. Proposing a speech model leads to characteristic

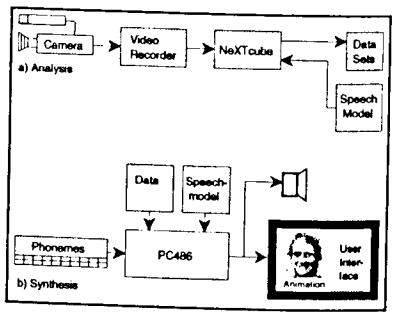


Figure 1. Visual speech analysis-synthesis system.

speaker-dependent movement data.The synthesis computer uses these data, together with the speech model, for the facial animation.

The data analysis consists of two steps, the fixing of the sound boundaries and the determination of phone-characteristic single pictures. Both steps have to be processed interactively.

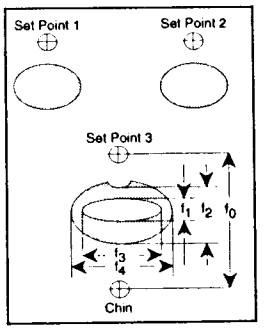
The phone boundaries were determined with the help of oscillogram, sonagram and acoustic feedback as described in [6]. Then, those video-frames fitting best with the subjective impressions of a well pronounced sound were indicated with the help of both the acoustic and visual material by different experts in lip-reading. In some cases, e.g. for the phonemes /h,g,k/, an exact determination was not possible; the production of these sounds is only weakly represented on the speaker's face and can usually not be perceived by people who lip-read. For this reason the existance of either one or none characteristic picture is proposed, concerning the modelling as well as the later facial animation [7].

The speech movements are characterized by continuous quasi-periodic opening and closing processes that are reflected in the courses of the visual features. Thus, the proposed motion model is based on a library of characteristic 'key-pictures', arranged in the extrema positions, that allow to track the courses of features. The interactively determined characteristic pictures of the phonemes are located in or at least close to these extrema.

In order to compose the library of key-pictures, the 'characteristic pictures' of the text corpus were classified by a FCM-algorithm (see [8]) with respect to lip shape and position by using specific visual features. The algorithm generates optimum location of the clusters automatically with respect to a given number of clusters. The cluster centers in the feature space compose the library of representative key- pictures.

In order to guarantee reproducable results some set points on nose and forehead and the contours of the lips were marked with a flourescent color. During the recording the persons were slightly radiated with UV light. An exemplary single frame shows Figure 2, together with the scheme of the feature extraction.

a



b

Figure 2. a. Examplary single frame, b. visual feature extraction of frontal view.

KEY-PICTURE SELECTION

A library of 'key-pictures' represents the properties of a speaker's speech movements. There is no a priori relation between the phonemes and the corresponding key-pictures. Thus, the unknown mapping function was trained in a complex artificial neural network with the help of the courses of visual features

for part of the spoken sentences. The other part is taken for the evaluation of the mapping quality.

The neural network considers forward and backward coarticulation effects by using an influence frame of 3+1+3 phonemes. The medium phoneme in its context is related to one (resp. none) key-picture. Pulling this 7×1-frame through a given sequence of phonemes creates a framework of pictures that represent important stages for the courses of visual features. The time distances between two key-pictures are calculated by averaging the time distances of the corresponding interactively determined pictures on the whole text corpus. The phoneme to be mapped is ignored if there was no characteristic picture to determine in the video film. A block diagram of the selection process is shown in Figure 3a.

Here, a key-picture out of the library is related to the phoneme /o:/ considering the dependencies of the preceeding and following phonemes /a:_n_h/ and /f_'_i/. The sign /'/ stands for an occlusional sound in a word boundary. The approximation of given facial movements for a sequence /Ph_1_Ph_2_..._Ph_8/ of phonemes by a sequence of key-pictures Kp_i is shown in Figure 3b. A morphing algorithm for grey-scale pictures creates interim frames at specific locations.

There are different approaches known for mapping phoneme sequences on speech pattern, as, for instance, the NETtalk algorithm [9]. This paper proposes the fuzzy neural network architecture seen in Figure 4 that considers similarities among the visual phonematic correlates. These serve for the input coding and, on the other hand, take influence on the network architecture [10].

For training, the FNN is cut at the viseme layer and trained in three steps by error-backpropagation:

- the phoneme-to-viseme-mapping with the help of the viseme structure of German (see [11]) and the recorded video material,
- the viseme-to-feature-vector mapping,
- the connected FNN with respect to the video material.

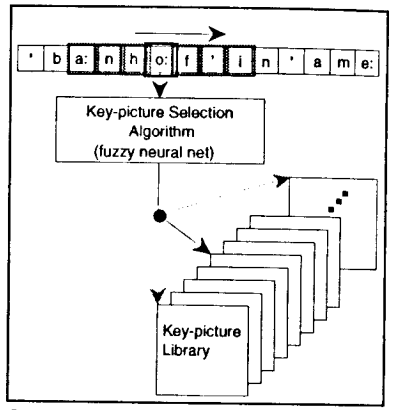


Figure 3. a. Selection, b. placement of key-pictures with respect to a given phoneme sequence Ph_1 ... Ph_8.

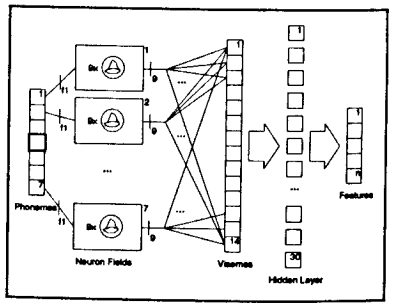


Figure 4. Block diagram of the fuzzy-neural network (FNN).

The exact placement of a key-picture in the frame of the phone boundaries is - depending on the context - also calculated with the help of an artificial neural network.

Evaluation of the Naturalness of the Facial Animation

For given phonetic input sequences, the resulting computer animation program produces similar courses of predicted visual features as measured in the video films. Despite of the fact that several artifacts still occur, the general tendency of opening and closing movements looks very much alike (see Figure 6).

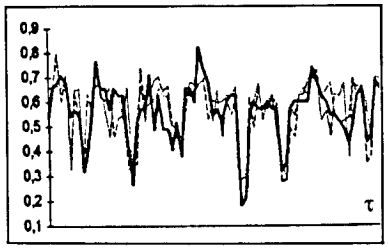


Figure 6. Measured (—) and predicted (- - -) courses of the visual feature $f_1(\tau)$.

Since even small local differences of the actual courses of visual features may result in a larger perceptual artefact and vise versa, the naturalness of the calculated videos has to be evaluated with the help of those who can lip-read, i.e. with hearing-impaired people. First results for a simple demonstration version of the animation program were investigated in a school for hard-of-hearing children and can be found in [12].

References

[1] Menzerath, P. and A. de Lacerda (1933), *Koartikulation, Steuerung und Lautabgrenzung*, Berlin.

[2] Fisher, C.G. (1968), "Confusions Among Visually Perceived Consonants", *J. Speech and Hearing Res.*, 11, 796-804.

[3] Storey, D. and Roberts, M. (1988), "Reading the Speech of Digital Lips: Motives and Methods for Audio-visual Speech Synthesis", *Visible Language* 22, 112-127.

[4] Cohen, M.M. and D.W. Massaro (1990), "Synthesis of Visible Speech", *Behaviour Research Methods, Instruments & Computers*, 260-263.

[5] H.H. Bothe, G. Lindner and F. Rieger (1993), "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, 45-49.

[6] C. Heise and H.H. Bothe (1993), "Phone and Syllable Segmentation by Concurrent Window Modules", *Proc. of the EUROSPEECH'93*, Berlin, 669-672.

[7] H.H. Bothe and F. Rieger (1994), "Zum Zusammenhang von akustischen und visuellen Korrelaten lautlicher Artikulationsprozesse", *Tagungsberichte der 20. Deutschen Jahrestagung für Akustik (DAGA)*, pp. 1337-1340, Dresden.

[8] Bezdek, J.C. (1981), *Pattern Recognition with Objective Function Algorithms*, London.

[9] Sejnowski, T.J. and C.R. Rosenberg (1986), "NETtalk: A Parallel Network that Learns to Read Aloud", *Electrical Engg. and Comp. Science Tech. Rep. JHU/EECS-86/01*, J. Hopkins University.

[10] H.H. Bothe (1995), "Fuzzy Input Coding for an Artificial Neural Network", *Proc. ACM Symposium on Applied Computing'95*, pp. 450-454, Nashville, USA, 1995.

[11] G. Alich (1961), *Zur Erkennung von Sprachgestalten beim Ablesen vom Munde*, (Doc. Thesis), Bonn.

[12] H.H. Bothe and F. Rieger (1993), "Computer Animation for Teaching Lip-reading", *Proc. 2nd European Conference on the Advancement of Rehabilitation Technology (ECART)*, pp. 4.4, Stockholm.