Neural Networks

A Visual Speech Model based on Fuzzy-neuro Methods

Hans H. Bothe

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

Abstract. This paper describes a new approach of modeling visual speech movements, based on a codebook of characteristic *key-pictures* and a complex fuzzy neural network (FNN). Goal is the development of a computer animation program as a training aid for learning lip-reading. The network architecture makes possible a fusion of linguistic expert knowledge into the FNN. 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 visual reflections of these movements on the speakers face may be seen as *visual speech.*

Whereas the smallest speaker-independent units derived from the acoustic signal being semantically distinguishable are the *phonemes*, there also exist smallest perceptible visual speech units , the *visemes*. The articulatory positions and transitions needed for the production of a phoneme find their visual expression in the related viseme. Knowledge about relationship and structure has been investigated in extensive experiments, especially with hearing-impaired persons. The key-work for German language is the dissertation of Alich [2]. According to these investigations, the 40 phonemes of German can be related to 12 visemes, whereas the phonemes /t, d, n/ and / , / have to be related to two different visemes, depending on the spoken context. Thus, for use of this knowledge in a computer-based facial animation system, the viseme scheme should be modified by isolation of those phonemes and defining two extra visemes. The following table 1 shows this modified scheme with eight consonant and six vowel visemes. The extra visemes V_{C5} and V_{V2} could be perceived as V_{C3}-V_{C5} or V_{V1}-V_{V3}, respectively.

Consonant Visemes				**Vowel Visemes**			
V_{C1}	/p, b, m/	V_{C5}	/l, r/	V_{V1}	/a, a: ë,/	V_{V5}	/o, œ/
V_{C2}	/f, v/	V_{C6}	/k, g, x, N, /	V_{V2}	/ä, Ä/	V_{V6}	/R/
V_{C3}	/s, z/	V_{C7}	/ç, j/	V_{V3}	/ i, i:, e:/		
V_{C4}	/t, d, n/	V_{C8}	/S , Z	V_{V4}	/o:, ø:, u, u:, y, y:/		

Table 1. Scheme of modified consonant and vowel visemes V_{C1}-V_{C8} and V_{V1}-V_{V6} of German.

Visual speech movements mostly contain sufficient information to enable hearing-impaired persons to lip-read a spoken text. Since the visual recognition is largely focussed on the speakers mouth region, especially on the lips, and the lip movements contain most of the visually perceptible information, this paper proposes the modeling of face movements based on the corresponding lip shapes.

2 Data Acquisition

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 multi-media 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 fluorescent color. To increase the contrast even more and to smooth the contours, the face was lightly UV-radiated so that shady parts of the lips as wrinkles or grooves were self-radiating instead of only reflecting light. The set points and contours were then localized with the help of an automatic contrast search program.

The two marked reference points and the set point on the nose refer to the head coordinate system and are used for correction of global head movements during the recording. This was necessary since an artificial fixing of the speakers head was felt to have influence on the naturalness of the articulatory movements. Three example frames of the video film, together with the feature extraction scheme, are shown in figure 1. Applying this scheme, each frame is represented by a five dimensional visual feature vector.

Fig.1. Characteristic video frames and extraction of primary visual features <m $_i$>.

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]. If in certain cases as, for instance, for the phonemes /h,g,k/, no characteristic frame could be determined, this information was used for the later facial animation.

The determined phone characteristic frames of the video film were classified in this feature space with respect to lip shape and position. The cluster centers compose a set of representative visual feature vectors and define a codebook of key-frames.

3 Generation of the Codebook of Key-frames

The feature vectors were classified with the help of of the fuzzy c-means algorithm as described in [3]. The iteration algorithm generates optimum location of the clusters automatically with respect to a given number of clusters. Thus, a desired number of key-frames can be defined by proposing the number of clusters before starting the classification process.

The fuzzy c-means clustering algorithm does iteratively calculate the cluster centers and with them the new membership grades of the objects. It may be processed mainly in the following four steps, with N being the number of sample vectors $\underline{m}_i$:

Step 1: Choose
- the desired number of clusters n with $2 \leq n \leq N$,
- the initial condition of the fuzzy membership matrix $\underline{\underline{U}}^{(0)} = (\mu_{ij})^{(0)}$ of the sample vectors $\underline{m}_i$ to the clusters Q_j.

Step 2: Adjust the n cluster centers $\underline{v}_j^{(s)}$ of iteration step (s).

Step 3: Calculate the membership values $(\mu_{ij})^{(s)}$ of the sample vectors for the new cluster centers.

Step 4: Stop iteration, if $\| \underline{\underline{U}}^{(s+1)} - \underline{\underline{U}}^{(s)} \| \leq \varepsilon$ with $\|...\|$ as a suitable matrix norm.

After the classification process, any straightforward phoneme-to-key-frame mapping is lost. Exemplary results with $n_1 = 15$ and $n_2 = 25$ clusters is shown in figure 2. The diagram draws the features width $\langle m_3 \rangle$ over height $\langle m_1 \rangle$ of the outer lip contour in arbitrary units, i.e. screen pixels. The centers of the circuits determine the cluster centers, and the radii the amount of sample vectors in that cluster.

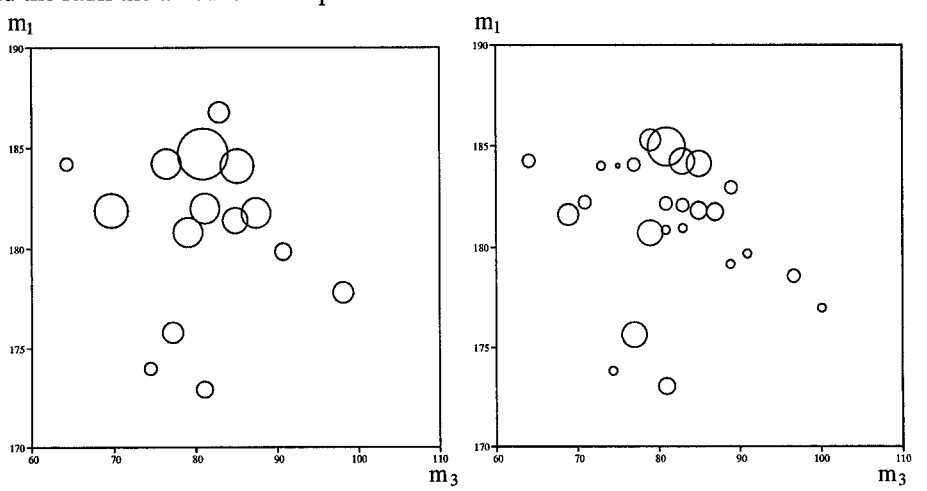

Fig.2. Exemplary classification results projected on a 2D space of the visual features $\langle m_1 \rangle$ over $\langle m_3 \rangle$ (units in screen pixels) with 15 or 25 clusters.

4 Key-frame Selection by a Fuzzy Neural Network

The subject of modeling visual speech and coarticulation effects has been addressed by several authors for different languages (e.g., [4-8]), whereas the movements are either directly controlled by certain visual features or calculated by a target and interpolation algorithm using a codebook of key-frames. The latter has the advantage of an easy implementation on computers with a relative small calculation power as, for instance, PCs; but a larger memory is needed for the storage of the codebook.

A first order approximation for modeling backward and forward coarticulatory effects takes into account the immediate next neighboring phonemes. For this purpose, the phonetic text can be split into overlapping diphones, diphthongs be represented by two closely connected single phones. The frame of the second is classified with respect to the first one [4]. This process leads to a deterministic diphone related phoneme-to-key-frame mapping. The frames may simply be depicted from a look-up table.

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. Thus, in this work a complex fuzzy neural network (FNN) was trained to map the given phoneme sequence onto a corresponding sequence of key-frames. It relates the single phonemes together with the surrounding next 3+3 neighbor phonemes to the frames of the codebook. For the training of the FNN, half of the text corpus was used, the other half being reserved for evaluation of the FNN generalization quality.

In the later speech synthesis, the same FNN is used for key-frame depiction, and the film is generated by calculating interim frames. The general design of the FNN is shown in figure 3.

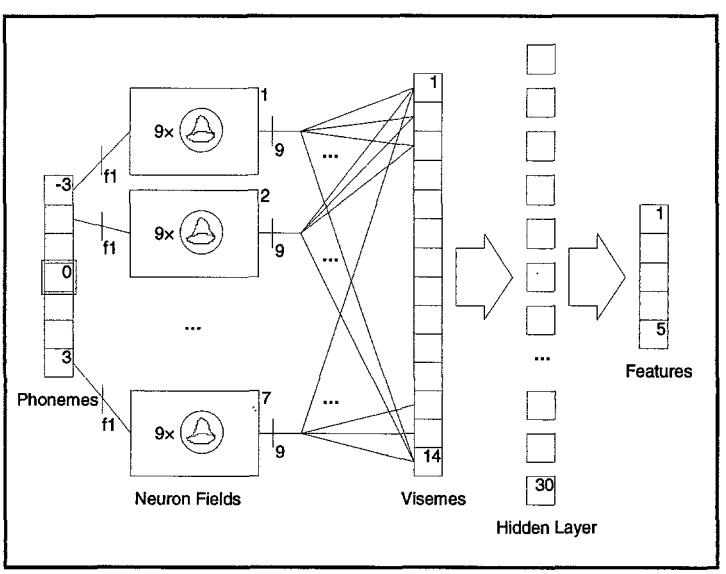

Fig. 3. Design of the ANN for key-frame selection.

The input coding is carried out by a self-organizing Kohonen map that represents the visual feature vectors correlated with the phonemes [9]. After the training process of the Kohonen map, each phoneme corresponds with a specific weight distribution over the map. The localization of the centers of gravity of these distributions, as shown in figure 4, may be taken as a complex similarity measure of the visually represented phonemes. The two dimensions x and y have no physical meaning, but rather serve for ordering the phonemes.

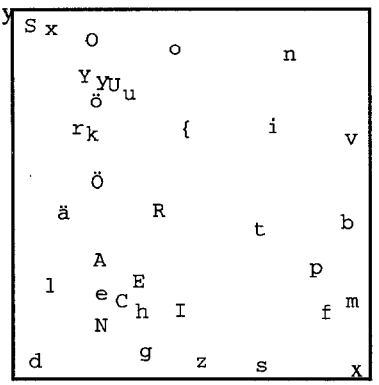

Fig. 4. Localization of the phonemes on a trained Kohonen map.

The FNN maps the input sequence of 7 phonemes - each out of the set of 41 phonemes - on a set of 14 output neurons by a radial basis function network (RBFN) with fixed Gaussian distribution functions. These overlapping functions cover the universe of the map. They are interpreted as fuzzy IF...THEN...-rules by detecting similarity of the phoneme locations. For instance, the phonemes /y, y:, u, u:, œ/ have about the same coarticulatory influence on the key-frame selection, but most probably a different one than /p, b, m, f/.

Each output neuron of the RBFN represents one viseme. This means that a certain phoneme in its context of 3+3 neighboring phonemes is represented by 14 membership values of the set of visemes. Thus, the existing linguistic knowledge on the perception of articulatory movements can be used in before adjusting the RBFN weights. The single neuron fields of the RBFN are interpreted as fuzzy rules, i.e. .

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 FNN with respect to the given phoneme-to-feature-vector mapping of the training sentences.

The FNN 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. The forecast of the short phoneme sequence /a:b/, together with the measured course of the features <m_1>-<m_4>, are shown in figure 5. The timing of the courses of features is calculated with the help of averaged key-frame distances.

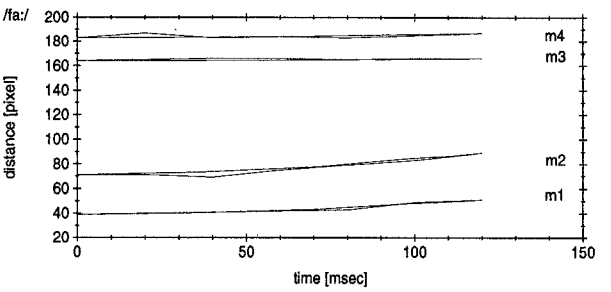

Fig. 5. Forecasted and measured courses of <m_1> to <m_4> for the phoneme sequence /a:b/.

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 N. v. Bötticher: Key-frame selection for the analysis of visual speech with fuzzy-c-means algorithm. In: B. Bouchon-Meunier & R. Yager & L.A. Zadeh (Eds.), Advances in Intelligent Computing, Springer-Verlag, Berlin-Heidelberg (to appear, 1995).
4. 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.
5. 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.
6. M.M. Cohen and D.W. Massaro: Synthesis of Visible Speech, Behaviour Research Methods, Instruments & Computers, (1990), 260-263.
7. 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 ESCA Workshop on Speech Synthesis, Autrance, (1990), 381-393.
8. F. Lavagetto: Converting Speech into Lip Movements: A Multimedia Telephone for Hard of Hearing People. Trans. Rehabilitation Engineering, (to appear; 1995).
9. Bothe, H.H.: Fuzzy input coding for an artificial neural network. (ACM / SAC'95), Nashville, (1995).