Startegies and Results for the Evaluation of the Naturalness of the LIPPS Facial Animation System

Jana Eger, Hans-Heinrich Bothe [1,2]

[1] Centre for Applied Hearing Research, Technical University of Denmark in Lyngby, Denmark
[2] Institute of Biomedical Engineering, Technical University of Berlin, Germany
hhb@elektro.dtu.dk

Abstract

The paper describes strategy and results for an evaluation of the naturalness of a facial animation system with the help of hearing-impaired persons. It shows perspectives for improvement of the facial animation model, independent on the animation model itself. The fundamental thesis of the evaluation is that the comparison of presented and perceived visual information has to be performed on base of the viseme structure of the language.

1. Facial Animation System

Speech-reading is an important skill for hearing-impaired persons to compensate for a loss of hearing. The larger this loss is, the more important becomes the additional visual detection of the correlating articulatory movements (*speech-reading of visual speech*). Especially in the case of a sudden loss of the ability to hear the visual speech perception has to be trained intensively for enabling continuous communication.

In order to gain quick learning effects, a facial animation system was implemented on PC with an open input vocabulary, which converts any given input text into corresponding articulatory facial movements. For the purpose of gaining a motion model, video films with prototype speakers were analyzed, taking the dependence of the movements on the complete spoken text into account.

The analysis system and functionality of the model-based animation has been described in several publications [e.g., 2-6].

As shown in fig. 1, the animation system uses phoneme sequences for conversion, and additional acoustic output can be gained with the help of a speech synthesis system [4]. The input phoneme sequences may be derived either by automatic text transcription (keyboard or ASCII file input), or with acoustic speech input.

The interface for phoneme surrender is based on probabilities for each phoneme (1…N) as given by the phoneme recognizer every 10 [ms]. Reasonable changes of the probability vector [ph_1 … ph_N] detect the phoneme boundaries, creating the natural dynamics of the facial animation.

The motion model is based on a set of 40 key-features, each of which consisting of inner/outer mouth contour and distance chin-nose.

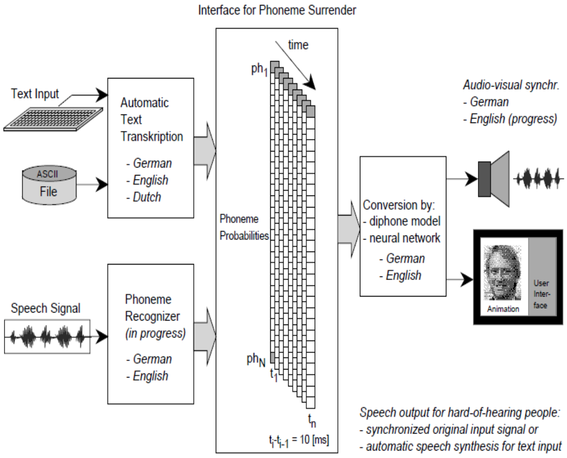


Figure 1: Training system for text or speech input and audio-visual output.

Depending on the given phoneme sequence, a background image is distorted to match the respective features. The position of teeth and tongue are later inserted.

Early versions of the animation system were cartoon faces with movable lips, teeth, tongue, eyes, and eye-brows. For the present evaluation, a 2D passport photo animation was used, in which a larger mouth and jaw region is animated between characteristic position vectors given by lips and jaw. Coarticulatory effects are taken into account by applying a diphone separation model for phonetic input sequences.

2. Evaluation Method

The exact process of visual speech perception is widely unknown, but the functionality is obvious. Whereas the *phonemes* are the smallest meaningful and speaker independent units of acoustic speech, equivalent units of the corresponding articulatory facial movements are the *visemes*.

2.1 Viseme Structure of the German Language

Facial movements which accompany phonemes belonging to the same viseme are perceived alike, e.g. /p,b,m/. The fundamental thesis of this paper is therefore that the evaluation of the visual information has to be performed with the help of the visemic structure of the language.

Consonant group of visemes		Vowel group of visemes	
B	/p, b, m/	A	/a, a:, R/
M	/m/ in final position	I	/ i, i:, {/
F	/f, v/	O	/o, o:, œ, ø:/
D	/s, z/	U	/u, u:, y, y:/
L	/l/	E	/e:, ä, Ä/
T	/t, d, n/		
C	/ç, j/		
G	/k, g, x, h, N, r/		
S	/S,Z/		

Table 1. Consonant and vowel visemes.

For German, the phonemes were grouped into 14 viseme classes, which are based on Alich's 1961 viseme model [1] and own investigations (table 1).

2.2 Realization of the Test

The test was carried out in three different schools for i) deaf and ii) hard-of-hearing children with 27 children of age 14-17. In each school, the full text corpus was presented in alternating order. The first and second day were reserved for program demonstration, and the following three days for the test realization. The presented text corpus consists of 68 German words, ordered in three lessons with i) short words (one to two syllables), ii) long words (four to six syllables), and iii) medium long logatomes (two to four syllables). The computer animation was based on correct phonetic input in order to avoid possible errors due to the phoneme recognizer; no additional acoustic signal was presented.

The short words were also used for getting easily into the subject of speech-reading from a computer screen, the longer words for covering contextual coarticulatory effects, the logatomes for exclusion of supplementary effects. Longer utterances were not used in order to avoid contextual supplements by the test persons. The corpus was constructed with respect to a balanced frequency of double visemes or *di-visemes* for German.

The test was carried out in four steps (fig. 2). In the first step the words were presented by teacher and computer (human facial movements and animation).

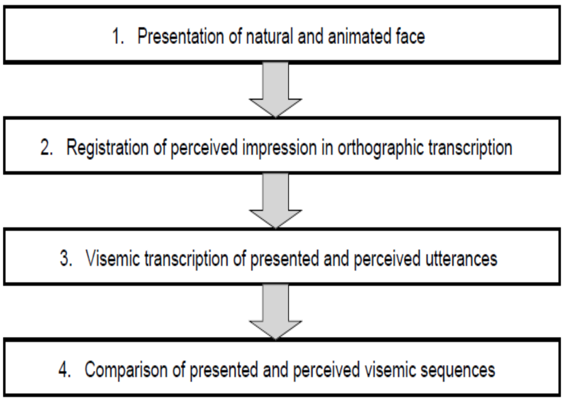

1. Presentation of natural and animated face
2. Registration of perceived impression in orthographic transcription
3. Visemic transcription of presented and perceived utterances
4. Comparison of presented and perceived visemic sequences

Figure 2. Block diagram for test and evaluation, consisting of four major steps.

The perceived *speech-read* information of both presentations was written down in orthographic transcription (letters) by the test persons and re-transcribed into phonetic sequences later by the test leader. This procedure is necessary since only few test persons have fundamental phonetic knowledge and the perceived information has to be fixed in written form.

For the comparison of presented and perceived information it is important that minimal pairs as, for instance, the words *Power-Bauer-Mauer,* are not distinguishable by the test persons, and a perceived *Mauer* for a presented *Bauer* cannot be treated as an error. Thus, the phonetic sequences have to be compared on a visemic basis, i.e. transcribed into viseme sequences before comparison. The following table 2 shows the visemic transcriptions for some exemplary words of the presented text corpus.

Word	Viseme sequence
Mutter	BUTA
Schüler	SULA
Postschalter	BOTSALTA
Puppenspieler	BUBITSBILA
Kippelfu	GIBILFU
Lollikopf	LOLIGOBF

Table 2. Visemic transcriptions for exemplary words.

3 Results and Interpretation

Three different statistics were applied to the results of the comparison. For a first overview, the recognition rates of human face and animation were compared with respect to different lessons (groups of words) and schools. Secondly, the correct, omitted, inserted, and mistaken visemes in initial, medium, and final position within the words were counted and compared for human facial movements and animation. Results are respective confusion matrices of single visemes. The third comparison is based on *di-visemes* in order to consider next-neighboring contextual influences or coarticulatory effects, respectively.

3.1 Comparison

The results of the first evaluation are shown in fig. 3.

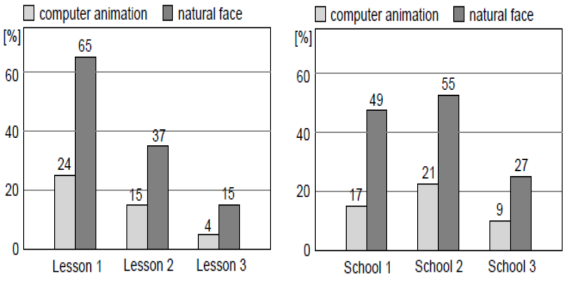


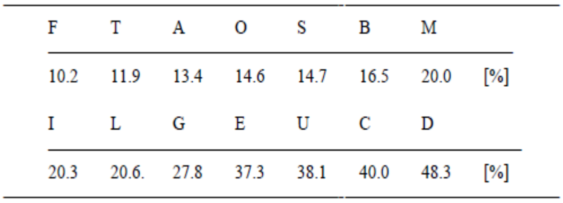
Figure 3. Recognition rates distinguished by lessons and schools: Lesson 1: short words, 2: longer words, 3: logatomes. Schools 1, 2: hard-of-hearing, 3: deaf children.

It can be seen that the short words were easiest to speech-read, followed by the longer words with a significant decrease of recognition rate. Logatomes showed relatively bad results for the computer animation and for the natural face, which clearly indicates the (desired) effect of lacking in content.

In the second diagram of fig. 3, the results schools for hard-of-hearing children (1, 2) are approximately within the same range, whereas those at the school for deaf children (3) were significantly lower. There might be two main reasons for that; at first, deaf people usually try to recognize complete gestures rather than partial actions (they show very bad results for the logatomes). Secondly, the knowledge of the language - also in written form - was by far lower for the deaf children. Whereas many could easily repeat the mouth movements, writing down the results in correct form turned out to become very difficult. For these reasons, the results lead to the conclusion that hard-of-hearing children do fit more into the proposed method of evaluation than deaf children. Further calculations were only taken from schools 1,2. When it had been recognized by the test leader that in some seldom cases a pupil refused to write down any text (which seemed to be too difficult at a whole), the respective words were ignored for both, the animation and the natural face.

3.2 Confusion matrices of single visemes

The second evaluation was performed in order to rank the quality of the single animated visemes. For this purpose, viseme confusion matrices were calculated for initial, medium, and final position. For demonstration, the values for all correctly recognized visemes were sampled for animation and natural face, independent on the position; the difference of which in [%] is shown in table 3. The values detect a ranking order of the animated viseme quality, taking the natural face as a reference.



F	T	A	O	S	B	M	
10.2	11.9	13.4	14.6	14.7	16.5	20.0	[%]
I	L	G	E	U	C	D	
20.3	20.6.	27.8	37.3	38.1	40.0	48.3	[%]

Table 3. Ranking order of the animated viseme quality.

The results show that (F, T, A, O, S) have a natural quality in most of the cases with a difference of animation and natural face of below 15 [%], followed by (B, M, I, L, G) of below 30 [%]. The differences for (E, U, C, and especially D) are so significant that they indicate a high necessity of correction for the animation model.

As one detailed result can be stated that a relatively high difference value was received for wrongly inserted (B) in initial position. Since most speech production processes are preceded by a longer phase of inhalation, a corresponding open mouth at the beginning of the animation has to be considered by the model; neglecting this simple fact will lead to the impression that the text sequence begins with a bilabial closure (B) ∊ /b,p,m/. A similar result holds for words with plosive sounds as /b,p,t/ in final position.

The high recognition error for (D) can at least partly be explained by wrong (vertical) positions of the tongue; this result was derived from a detailed discussion started after calculation of the above error rate.

Additionally, an artificially inserted error could easily be detected. The viseme (U) was relatively often confused with the (O), which is due to the fact that the animation uses the same key-pictures for (U) and (O, independent on contextual coarticulatory effects).

3.3 Confusion matrices of di-visemes

The used conversion model for the animation system is based on key-picture selection with the help of a di-viseme model [see 2-6] in order to take contextual influences into account. For evaluation of the quality of picture selection, respective confusion matrices for di-visemes were calculated for natural and animated facial movements.

As a result, the most significant differences occurred for the di-visemes (UD, EU, DA, DO, DU). Again, this can be explained with wrong tongue positions for (D) which is obviously visible for these transitions. The involved vowels do not have a forced tongue position by themselves, and coarticulatory effects must be taken into account also for articulatory tongue movements.

Conclusion

The designed test strategy for subjective evaluation of the animation system by visual human perception allows to rank the presentation quality and to detect artifact facial movements. Precise and specific information about necessary improvement of the animation can be gained, taking the results of the human face as reference values into account.

References

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

[2] Bothe, H.-H., F. Rieger (1992): Lipreading - analysis and synthesis on microcomputers. In: W. Zagler, Proceedings of the International Conference on Computers for Handicapped Persons ’92, Vienna, R. Oldenbourg, Vienna-Munich.

[3] Bothe, H.-H., G. Lindner, 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 (Proceedings of the 1st TIDE Congress, Brussels), IOS Press, Amsterdam.

[4] Bothe, H.-H., E.A. Wieden (1994): Artificial visual speech, synchronized with a speech synthesis system. In: W.L. Zagler, G. Busby und R.R. Wagner (Eds.), Lecture Notes in Computer Science, Vol. 860, Springer-Verlag.

[5] Bothe, H.-H. (1995): Artificial Visual Speech, Generated by Fuzzy Inference Methods. In: I. Placencia-Porrero and R. Puig de la Bellacasa (Eds.), The European Context for Assistive Technology (Proceedings of the 2nd TIDE Congress, Paris), IOS Press, Amsterdam.

[6] Bothe, H.-H. (1995): Relations Between Visible and Audible Speech Signals in a Physical Feature Space: Implications for the Hearing-impaired (Invited Lecture). NATO Advanced Study Institute ’Speech-reading by Man and Machine: Models, Systems and Applications’, Chateau de Bonas, France.