Feasibility of Face Animation on Mobile Phones for Deaf Users

Takács György, Tihanyi Attila, Bárdi Tamás, Feldhoffer Gergely,

Pázmány Péter Catholic University Faculty of Information Technology
1083 Budapest Práter utca 50/a
Telephone: +(36) 18864763, Fax: +(36) 18864724,
email {takacs.gyorgy, tihanyia, bardi, flugi,}@itk.ppke.hu

Abstract **Mobile telephone systems have made a fantastic progress in general, but deaf people are practically excluded from the benefits of mobiles. Our aim was to develop a new communication aid for deaf users. Our system directly converts the audio speech signal into the video of animated face, so the deaf users can receive voice messages by lip-reading. Our system was implemented and tested in a PC environment earlier. This paper reports on the implementation on mobile phones and PDAs. The implementation problems and the potential steps for further improvements are also discussed.**

I. INTRODUCTION

The information in speech is conveyed not by pure audio signal but mixed audio and video signal for deaf and hard of hearing people. Visual signal of lip movement is a partial but very important basis for communication with the hearing society for deaf people. [5;6;7;8] Mobile phone systems have changed radically the telephone communication but deaf people are still excluded from this. The 3rd generation mobile systems support the real time video communication but the everyday application is still not typical. Real benefits for deaf people can start only if the penetration of 3G sets reach a minimum of 80% and the tariffs for real time broadband communications will be affordable for deaf people.

Our new solutions try to provide communication aids for deaf users on the basis of cheap 2G networks and terminals.

Nowadays the processing power and memory capacity of mobile sets and GSM capable PDAs are relatively high compared to older types but different class as desktop PCs. New operation systems (Windows Mobile and Symbian) compete for a higher marker share.

Our new applications calculate with the well developed practice of deaf persons in the field of lip reading. Our communication aid converts the speech audio signal into video signal of animated speaking face.

II. SYSTEM DESCRIPTION

Our complete system is able to convert the audio speech signal into video signal of the animated speaking face. Deaf users can understand the speech message based on the speaking face video. The elements of the system are described below. These are software implementations applied on normal mobile phone.
The conversion process converts the analogue speech signal into a series of video frames of an animated face. The first step is a conventional feature extraction from the speech signal. 16 MFCC coefficients are calculated frame by frame.

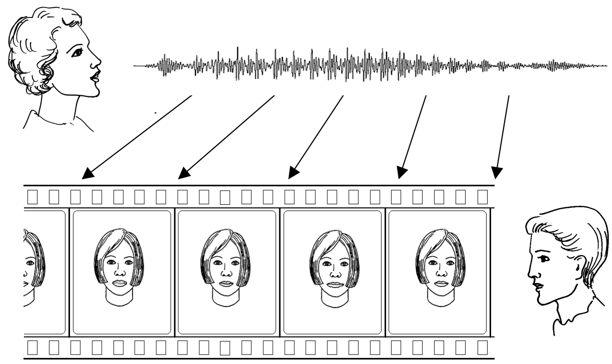

Fig. 1. The main process of the signal conversion

The coefficients from 5 successive frames compose the input of the artificial neural network. The outputs of the neural net provide the 6 compressed PCA parameters to describe the actual shape of the mouth on the speaking face. The speaking face is represented by an MPEG-4 standard based feature point coordinate set. [1;2;3;11] The 15 element subset of the original 84 feature point set is used in our model to animate the mouth and its environment on the speaking face. Other feature points like the representation of the eye and the nose are fixed because their role in lip reading has limited importance. The 6 PCA components can represent the 15 feature points with a maximum of 2% error. The training of the neural net was a critical element in the system development. The training material was carefully selected from the audio and video records of professional lip-speakers. Further details of the system can be found in [12;13;14].

III. "2D OR 3D" REPRESENTING OF THE ANIMATED SPEAKING FACE

The importance of 3-dimensional visual information in lip-reading was carefully tested during the system development. This test was designed to decide whether to use 2D or 3D motion tracking in our audiovisual database. In 2D motion tracking, horizontal (x) and vertical (y) coordinates of the object are captured, regardless of its depth (z) coordinate. In 3D motion tracking x, y and z coordinates are captured with multiple camera systems. The advantage of 3D motion tracking is obviously the depth information. Its disadvantage is the costliness, and the procedure to eliminate inconsistencies of images coming from different cameras

makes additional inaccuracies in determination of x and y. This experiment helped us to rate advantages and disadvantages in respect to lip-reading.
Generally, the human vision experience of depth in the 3-dimensional space is a result of a very complex reconstruction process in the brain. Several methods can be activated and integrated in this process based on different visual cues of the primarily 2-dimensional images coming from the eyes.
Watching videos (normal 2D moving pictures), only dynamic and pictorial information can count in one's depth perception. In our test videos, the lower half of the speaker's face was shown in frontal view, as it can be seen in Figure 2. In that sort of videos, the leading cues of depth perception are represented with different colours (pictorial), namely shading effects and texture. These cues can be almost completely hidden by distorting images. A simple transformation, a threshold based binarization was applied to the original videos. Each pixel of the images was set to white or black if its brightness exceeded a threshold or not. See Figure 2/b. Due to that distortion operation, the shape of the mouth can be tracked on the xy-plane well, but the feeling of spatial forms relies on former experience, not on real 3D information.

2/a: Shot from original video.

2/b: Shot from binarized video.
Fig 2: Example shots from videos of "2D or 3D" test.

Original and binarized videos were shown to deaf test persons (N=8). 24 short clips were shown with no sound. The task was to recognize what two digit number was said by the speaker. The speaker was a teacher of a secondary school for deaf students; her articulation was clear and favourable to lip-reading. The clips were displayed on the screen of SonyEricsson P910 smart-phones (40*61 mm). The phones were handheld by the test persons. Videos were compressed to 3gp file format at 15 frames per second, 176*144 pixel resolution, and 240 kilobits per second.
Though we expected the falling of the correct recognition rate in the case of binarized videos, surprisingly there was not significant difference. The result of this experiment was 83.0% correct recognition for videos with natural colours and 82.1% for distorted videos. We think that this difference is too small to be meaningful beside the limited number of subjects and task items. Consequently, we decided to use only 2D motion tracking in our database and in the face model.

IV. SYSTEM IMPLEMENTATION ON WINDOWS MOBILE BASED PDA

The first implementation of our system used a normal desktop PC. The available resources on this PC were enough for real time implementation of the system including the extracting the MFCC coefficients from the speech signal, calculating the PCA parameters by neural network, calculating the feature point coordinates and running the MPEG-4 face model.
The present PDAs have processors with ARM architecture instead of traditional Intel processors, 240x320 pixel portrait shape display instead of high resolution landscape one, Windows Mobile operation system instead of standard Windows, touch-screen input instead of keyboard and mouse. The PDA implementation has to focus on limited memory and energy.
The real problems in the software design were related rather to memory usage and energy usage then the implementation of tested algorithms.
Using Windows Mobile each display representation creates a window class. Such a product is a template containing several widow properties and procedures executed during different window functions. The most important among them is the window procedure describing the execution of instruction handling. Messages can be received from the operation system, or from another window. The message contains information and commands for our window and the window has to react on it. E.g. message might be moving or resizing the window or a keyboard input. The Windows Mobile handles such messages organised as a series of windows. The program designers decide which messages will be handled or ignored.

V. IMPLEMENTATION OF MOVING MOUTH BY POLYGON

The moving mouth was implemented by a polygon as the simplest solution. The corner points of the polygon are equal to the 8 feature points of the MPEG-4 face model representing the mouth. The polygon is drawn by a

DoPaintMain predefined procedure. The actual parameters of the mouth are defined by a matrix calculated according to the system elements description described above. The size of the matrix is 16x291 because the 8 feature points mean 16 coordinate values and the utterance consists of 291 time frames. The time frame size is 40 ms. The polygon interconnects the feature points by a straight line. The resources in the PDA are enough for further refinement of the animated mouth. Such procedure will be described in the next point.

VI. BACKGROUND FACE PICTURE

The moving polygon shape mouth model has no uncoloured background but a picture of a photorealistic face. Frame by frame the part of the background picture surrounding of the mouth is recalculated. The original background picture is loaded by SHLoadDIBitmap instruction into the Device Context. So we can copy the replica of the display in the memory. In the implementation of face animation frame by frame the actual face picture is calculated from this memory content.

The MPEG-4 standard defines 5 distances to describe the main parameters of the face (like distance of eyes, width of the mouth…). These distances can be used to recalculate the feature point coordinates using different size or shape of the reference face or different movement amplitude. So the background face picture can be changed after distance based recalculation.

Further enhancement of the polygon shape mouth movement on the background face picture can be implemented by morphing of individual points on the background picture.

VII. MORPHING

In the beginning solution the mouth was drawn by a polygon. The problem of this simple method referring to coordinates outside the point is difficult. So drawing the mouth is possible by points and to adjust the colour of the points needs manual interaction. In this phase the points within 5 pixel distance from feature points have temporarily black colour (the final colour will be assigned in a later phase). This calculation has a simple algorithm but it uses the processor intensively. The colour assignment steps use SetPixel instruction. The calculation processes run in the background, the refreshment of the display is executed after the calculations.

The morphing in our context means that the mouth and the surrounding points have seamless transitions from one frame to the next. [4;9;10;15] Objects in the source picture have seamless transition to the objects of the target picture. Higher distance from the feature point means lower movement of the point. The movement is calculated by formula cos(a/N*3.14)+1 where a is the distance from the feature point and N is the radius within we calculate with movement.

The actual colour of a point is calculated from the colour of the nearest two feature points. The starting (natural) position and the colour of feature points are known. The colour of a moved feature point remains the same. The colour calculation of non-feature points is more complicated. First we calculate where the actual point is from and its colour equals with the colour of the equivalent point in the starting (natural) picture. The colour assignment steps use SetPixel and GetPixel instruction.

The morphing procedure is a very processor and memory intensive one.

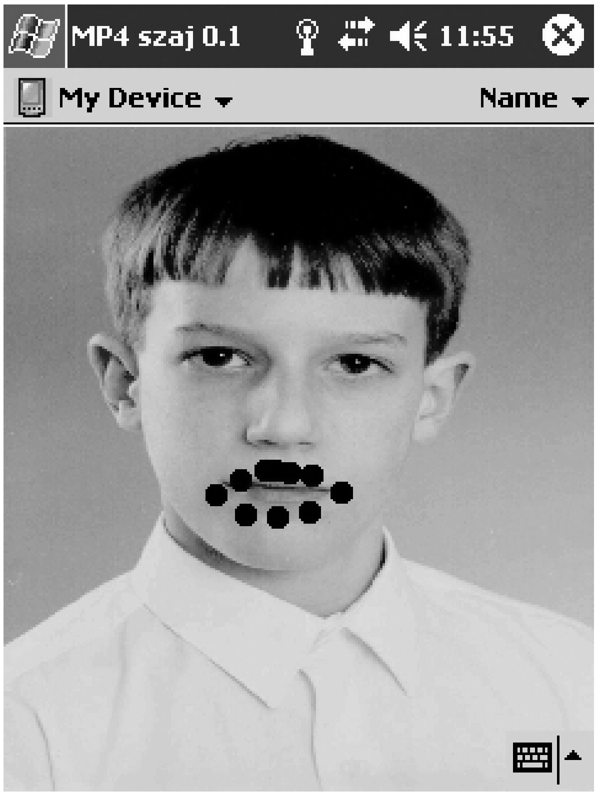

Fig 3. The centre of dots represent the centre of morphing. The area of dots cover the points which are moving during morphing.

One really problematic area in the model describer above the handling of neighbour points which move independently and violate the morphing rule. One example for this the point pair: the first belongs to the upper lip and the second to the lower lip in the case of a closed mouth. By opening the mouth these points are separated and not related by the morphing rule. Border lines can be defined to describe special areas where the morphing rules are not valid

VIII. DISCUSSION

The example implementation has been successful using MDA-III type PDA. The most significant limitation is the animation of the mouth by only 8 feature points. The preliminary tests indicated that 8 feature points are able to transmit the most important speech information on the animated face.

The more developed processes in morphing of the points of the mouth can reduce the processing demand of the animation so the available processing capacity can be used to increase the number of the feature points. More feature points can add more details to the picture, so the naturalness and readability of the animated face can also be improved.

The quality of the PC based implementation was carefully tested [12]. The principles and the main technical parameters are the same at the PDA implementation so the results of earlier tests are also valid. The deaf users were able to recognise correctly 48% of the independent isolated words based on pure lip reading. The animated face model was controlled by feature point parameters calculated directly from the speech signal. Deaf people have enhanced abilities to understand complete speech messages based on partially recognised elements based on the redundancy of the human speech. So the 48% value in the recognition of independent words is an acceptable level for communication aid for deaf people.

IX. CONCLUSION

Our experiments have proved that the PC tested voice to facial animation conversion can be implemented in mobile phones or in GSM capable PDAs. Matching with the limited available resources need some non-conventional program design solutions.

Life demo on the conference will be presented.

Introduction of more feature points in the model like at the internal contour of the mouth can improve the quality of the system according to the feedback from deaf testers.

X. REFERENCES:

[1] Balci, K. Xface: MPEG-4 based open source toolkit for 3D facial animation. In Proc. Advance Visual Interface, 2004.

[2] J. Beskow, Talking Heads, Models and Applications for Multimodal Speech Synthesis, Doctoral Dissertation, KTH Stockholm, 2003.

[3] P. Cosi, A. Fusaro, G. Tisato, “LUCIA a New Italian Talking-Head Based on a Modified Cohen-Massaro’s Labial Coarticulation Model”, Proc. of Eurospeech 2003, Geneva, Switzerland, pp. 2269-2272, September 2003.

[4] Garofolo, J., Lamel, L.F., Fisher, W.M., Fiscus, J.G., Pellet, D.S., Dahlgren, N.L, 1988. The DARPA TIMIT CDROM. Available from LDC: http://www.ldc.upenn.edu/cgibin/aesl/aesl

[5] B. Granström, I. Karlsson, K-E Spens: „SYNFACE – a project presentation” TMH-QPSR - Fonetik, vol. 44, pp. 93-96, 2002.

[6] R. Gutierrez-Osuna, P. K. Kakumanu, A. Esposito, O. N. Garcia, A. Bojorquez, J. L. Castillo, I. Rudomin, “Speech-driven Facial Animation with Realistic Dynamics” IEEE Transactions on Multimedia, vol. 7. pp. 33-42, February 2005.

[7] M. Johansson, M. Blomberg, K. Elenius, L.E.Hoffsten, A. Torberger, “Phoneme recognition for the hearing impaired” TMH-QPSR - Fonetik, vol. 44, pp. 109-112, 2002.

[8] P. K. Kakumanu, A. Esposito, O. N. Garcia, R. Gutierrez-Osuna: „A comparison of acoustic for speech-driven facial animation“ Speech Communication 48 pp 598-615, 2006.

[9] S. E. Palmer: Perceiving surfaces oriented in depth. In: Vision Science: Photons to Phenomenology, MIT Press, 1999.

[10] I. S. Pandzic, R. Forchheimer (Eds.), MPEG-4 Facial Animation: The Standard, Implementation and Applications, Wiley, Chichester, 2002.

[11] Pelachaud C., Marno Caldognetto E., Zmarich C., Cosi P., “Modelling an Italian Talkting Head”, in Proceedings of AVSP 2001, Aal-borg, Denmark, Septembre 7-9 2001, pp 72-77

[12] G. Takács, A. Tihanyi, T. Bárdi, G. Feldhoffer, B. Srancsik: “Speech to facial animation conversion for deaf applications” 14th European Signal Processing Conf., Florence, Italy, September 2006.

[13] G. Takács, A. Tihanyi, T. Bárdi, G. Feldhoffer, B. Srancsik: “Database Construction for Speech to Lip-readable Animation Conversion” 48th Int. Symp. ELMAR-2006 on Multimedia Signal Processing and Communications, Zadar, Croatia, June 2006.

[14] G. Takács, A. Tihanyi, T. Bárdi, G. Feldhoffer, B. Srancsik: “Signal Conversion from Natural Audio Speech to Synthetic Visible Speech” Int. Conf. on Signals and Electronic Systems, Lodz, Poland, September 2006.

[15] G. Salvi: ”Truncation error and dynamics in very low latency phonetic recognition” Proc. Of ISCA Workshop on Non-linear Speech Processing, 2003.