Mobile SynFace:
Ubiquitous visual interface for mobile VoIP telephone calls

Fernando López-Colino
UAM
Department of Computer Science
Madrid, Spain
fj.lopez@uam.es

Jonas Beskow
KTH
Speech Music and Hearing
Stockholm, Sweden
beskow@kth.se

José Colás
UAM
Department of Computer Science
Madrid, Spain
jose,colas@uam.es

Abstract

This paper presents the first version of the Mobile Synface application, which aims to provide a multimodal interface for telephone calls on mobile devices. The Mobile Synface application uses a talking face; it will simulate realistic lip movement for incoming voices. This application works as a complement for mobile VoIP applications without modifying their code or their functionality. The main purpose of this application is to improve the usability of mobile voice communication for hard of hearing people, or in noisy environments.

1 Introduction

Hard of hearing people are handicapped in mobile telephone use because of the lack of visual information; what's more, mobile telephones are usually used in noisy environments and this makes communication even more difficult. SynFace (Beskow et al., 2004) attempts to provide a solution to the problem by providing visual speech support for telephone calls. The interface is based on a talking head which speaks the callers' words. In order to achieve this feature, the speech is recognized at the phoneme level, and the corresponding facial animation parameters are generated.

The *Hearing at Home* project (Beskow et al, forthcoming) focuses on the needs of hearing-impaired people in home environments. The project is researching and developing innovative solutions to support perception of speech and audio in domestic environments.

Nowadays, wired communications are being replaced by wireless solutions. People have wireless phone terminals in their houses so they can move freely while talking on the telephone. The proliferation of mobile phones shows that people do not want to be restricted to a fixed place when talking on the telephone. The Synface project was developed for desktop computers, so any person using the application cannot move freely because they have to be in front of their computer screen. The Mobile Synface application aims eliminate this disadvantage.

2 Description of the Application

The desktop version of the SynFace application manages the whole process including speech recognition, visible articulation generation and face rendering. Mobile devices today cannot manage all the required processes, so the solution implied a distributed architecture limiting mobile device processes to the essential ones. For this reason, only the face rendering part may be performed on the PDA (see Gjermani, 2008 for details), while speech recognition and visible articulation generation should be performed on a server.

The new requirements and limitations implied an important change in the systems architecture. As was stated before, the speech recognition engine is located on the server, so voice should be routed through it. This new requirement implied the development of a conference server which allowed managing both VoIP calls, the call between the remote caller and the server, and the call between the server and the mobile device. The VoIP conference server was developed using the PJSIP project[1], an open source SIP client/server software. This server receives an incoming call, and after creating a new call to the mobile device, it connects both of them. When the connection is successful, the incoming audio from the first user, is also redirected to the speech recognizer.

One of the main objectives of the application is to be completely independent of the VoIP application. This is an advantage because the Mobile Synface application would work with most of VoIP applications as long as the VoIP confer-

[1] http://www.pjsip.org/

ence server is prepared to manage their protocols. But it also implies that voice-lip synchronization cannot use any information provided by those applications. For this reason, a synchronization protocol has been introduced between the mobile application and the server. In addition to this, the mobile application estimates the network delay for every frame request. Using this data, the application tries to compensate this delay in order to maximize speech-lip synchrony. The resulting relation between lip movement and speech is shown in Image 1, when comparing audio wave and the jaw opening parameter.

The communication between the mobile device and the server uses both TCP and UDP protocols. Each rendered frame is defined by corresponding animation parameters provided by the server. That information is generated in real time, so when a frame is going to be rendered that information should be requested. In order to minimize the network delay in the request-response process short UDP packages are used for this task. Only synchronization commands, because their critical importance, use TCP messages. This protocol only requires a low bandwidth, approximate value of 2kbps, so the VoIP application can use most of available bandwidth.

The application is resource demanding for the mobile device. For the first development the chosen device was a Dell Axim X51v PDA, running Windows Mobile 5.0. It is equipped with an Intel PXA270 624MHz CPU, and an Intel 2700G graphics chip, with hardware supported 3D rendering via the OpenGL ES API[2].

As it has been stated before, rendering time depends on network delay and mobile device hardware resources. Also, the third party VoIP application requires network and processing resources, so the Mobile Synface application performance deteriorates.

The mobile version of the PJSIP project SIP agent has been used for call management. This simple application manages VoIP calls using the SIP protocol. In order to improve global performance and decreasing network usage, the VoIP application was configured with a low audio quality setting, audio was sampled at an 8 kHz rate. The codec chosen for the test was the G.722 (PCMU), which provided good balance between CPU usage and network load.

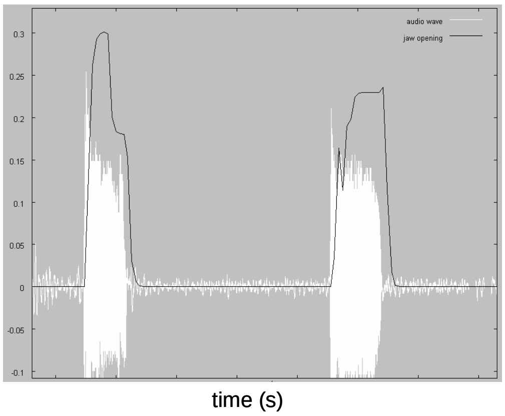

Image 1: Voice-lip synchronization at the mobile device – the speech wave is shown together with the jaw opening parameter.

3 Evaluation

Evaluations to date have been laboratory based, and focused on objective measurements of the application's performance. These experiments have been focused on testing voice and face synchronization, and measuring animation frame rate. The obtained results show that the delay between speech wave and face movement is lower than 90 ms, which is within acceptable values. The observed frame rate when the complete system is working (an average value of 9-10 fps) does not provide a desirable smooth animation. However, the final result does not show flickering.

Acknowledgments

Authors would like to appreciate the FPU-UAM research fellowship and for its financial support and the Hearing at Home project (EU IST-045089).

References

Beskow, J., Karlsson, I., Kewley, J., & Salvi, G. (2004). SYNFACE - A talking head telephone for the hearing-impaired. In Miesenberger, K., Klaus, J., Zagler, W., & Burger, D. (Eds.), Computers Helping People with Special Needs (pp. 1178-1186). Springer-Verlag.

Beskow, J., Granström, B., Nordqvist, P., Al Moubayed S., ., Salvi, G., Herzke, T., & Schulz, A. (forthcoming). Hearing at Home – Communication support in home environments for hearing impaired persons. In Proceedings of Interspeech 2008. Brisbane, Australia.

Gjermani, T. (2008). Integration of an Animated Talking Face Model in a Portable Device for Multimodal Speech Synthesis. Master thesis, KTH.

[2] http://www.khronos.org/opengles/