TOWARDS A HIGH QUALITY FINNISH TALKING HEAD

Jean-Luc Olivès, Mikko Sams, Janne Kulju and Otto Seppälä
Helsinki University of Technology
Laboratory of Computational Engineering, P.O. Box 9400, Fin-02015 HUT, Finland
Matti Karjalainen, Toomas Altosaar, Sami Lemmetty, Kristian Töyrä and Martti Vainio
Helsinki University of Technology
Laboratory of Acoustics and Audio Signal Processing, P.O Box 3000
Fin-02015 HUT, Finland
Email: {first name}.{last name}@hut.fi

Abstract: We described here how our Finnish talking head will be improved by using a new auditory speech synthesis method based on neural networks and an optimal synchronization of the facial speech animation and the audio.

INTRODUCTION

In face-to-face communication speech perception is based on both visual and auditory information. In addition to providing us with non-linguistic information of e.g., the talker's identity, age, emotions, and spatial location, visual speech also increases the intelligibility of the message. Contribution of visual speech becomes particularly prominent when the auditory speech is degraded due to noise [1] or hearing impairment [2]. Although not as effective as natural faces, synthetic faces also increase the intelligibility of both natural and synthetic auditory speech [3, 4].

Audio-visual speech synthesizers, 'talking heads', are being developed in many laboratories around the world [5]. Numerous applications will benefit of high-quality visual speech synthesis including human-computer interfaces, video coding (MPEG4 SNHC [6]), and communication aids for hearing impaired people.

We have developed a first version of a Finnish talking head [4] in which the user types in text and has both synthesized auditory speech and synchronized facial animation are created automatically. We have combined a 3D facial model with a commercial auditory text-to-speech synthetizer (TTS). The auditory speech is produced by concatenating pre-recorded samples of natural speech according to a set of rules [7]. Sample animations can be downloaded from our web site [8].

The quality of the current speech synthesis is not yet adequate. A new strategy has been developed to improve the TTS and to integrate the auditory synthesizer with facial animation. Our present synthesizer suffers from inaccurate synchronization, especially when the hardware capabilities are limited. We are developing a new method to achieve an optimal synchronization, independent of the platform used. This method is based on predictive visual synthesis.

0-7803-5610-1/99/$10.00 © 1999

ACOUSTIC SPEECH SYNTHESIS

The strategy of developing a new high-quality text-to-speech (TTS) synthesizer for Finnish [9] is based on warped linear prediction for voice synthesis and two alternative strategies for controlling this synthesis process: neural networks or parameter codebook lookup in addition to a set of rules. A system level architecture of the current prototype is shown in Figure 1.

Warped linear prediction (WLP) [10,11] is a version of linear prediction that is based on bilinear conformal mapping of unit delay to implement digital filters on a warped frequency scale. A particularly interesting case is to warp the frequency so that it corresponds to the psychoacoustically motivated Bark scale [12]. This means that the frequency resolution of warped algorithms corresponds closely to the resolution of human auditory perception.

WLP makes it possible to reduce the synthesis filter order remarkably when wide-band audio signals are processed: in our case with a sampling rate of 22 kHz a suitable WLP order is 12, compared to filter order of about 24 when ordinary linear prediction is used.

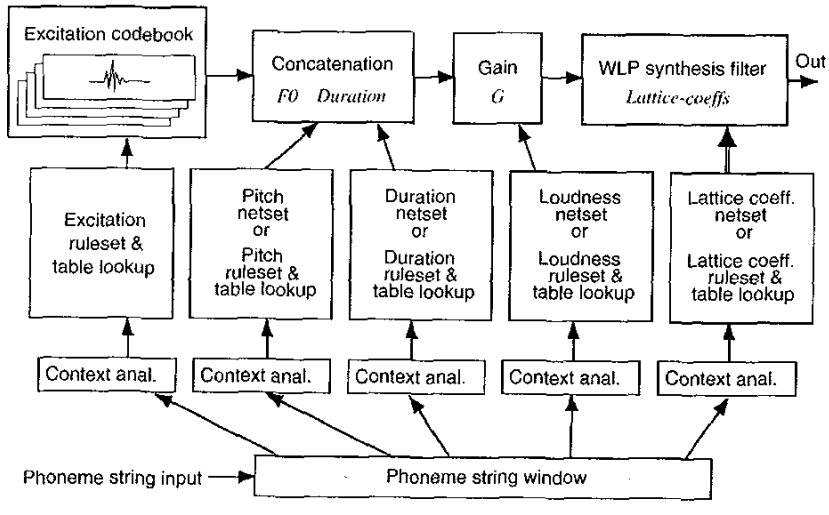

Figure 1: Block diagram of the voice synthesis

The synthesis of voice signal (top row blocks in Figure 1) is carried out by concatenating WLP inverse filtered residual signals from an excitation codebook, and controlling pitch (F0) and amplitude (G). This is then filtered by WLP synthesis filter to yield the synthetic speech output. The parameters of the synthesis process are updated in the beginning of each new pitch period. The excitation codebook contains context-dependent inverse filtered pitch-sized signals analyzed for a specific speaker from a speech database.

The control parameters are computed with different methods in the two alternative strategies. Using multilayer perceptron nets (MLP) allows a memory-compact realization where numerically coded features of phonemes in a context frame (3

previous ones, the current one, and 3 next ones) are composed into a context vector that is input to the MLP nets. Neural networks are specialized so that a set of nets is used and a proper one is selected based on which phoneme class, for example, is to be synthesized. The neural networks are trained with data analyzed from the same speech database entries that are used for excitation analysis.
Another strategy with higher speech quality is based on control parameter codebooks. The entries of the training speech database are analyzed in order to compile the control parameters and excitations for a rich set of phonemic contexts so that during synthesis the context of most resemblance to the one to be synthesized is searched for and the parameters are used for WLP synthesis. The excitations and WLP lattice parameters can be linearly interpolated in time over relatively long segments. With this strategy the tone quality is very high and mimics much the individual human speaker who had spoken the speech database items. This realization, however, needs more memory than the neural net approach.

VISUAL SPEECH SYNTHESIS

Visual speech synthesis is based on a letter-to-viseme mapping. A viseme is a category of similar visual speech articulations. For example the phonemes /b/, /p/ and /m/ have the same viseme (mouth closed). In this approach, each letter of a written text corresponds to a viseme. The derivation from letters instead of phonemes is justified in the Finnish language, because in Finnish there is a very strong link between letters and phonemes.

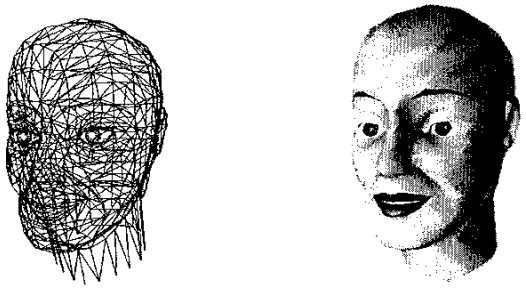

Figure 2: Wireframe and normal view of our talking head.

Our facial model (see Figure 2), a descendant of Parke's model [13], is a parameter-controlled topological model. Ears, teeth and the back of the head have been added to improve the naturalness of the model. It is currently controlled using a set of 49 parameters. We have added four new parameters to the original model, in order to improve control near the mouth region.
Twelve parameters affecting jaw rotation and lip shape are used for visual speech. These parameter values were determined simply by changing them so that the shape of the head's mouth matched that of the programmer's mouth as seen in a mirror. Visual speech is animated by linear interpolation between visemes.

SYNCHRONIZATION OF THE TALKING HEAD

The relationships between the component parts of the audiovisual synthesizer are presented in the Figure 3. The text input is transmitted to the audio I/O queue. The function of this queue is firstly to provide the audio synthesizer with the text, and secondly to buffer the sound blocks produced by the synthesizer. When the audio queue contains a full diphone, it's duration and the corresponding text is sent to the video queue. From the text information, we get the two visemes (the start and the end position) which have their own facial parameter sets. The animation is created between these sets. The number of frames between the visemes is defined by the ratio of the diphone duration and the frame rate. The parameter sets needed for producing these frames are interpolated between the two extreme facial positions. The video queue is filled with these parameter sets and the first image is synthesized, although it is not shown yet.

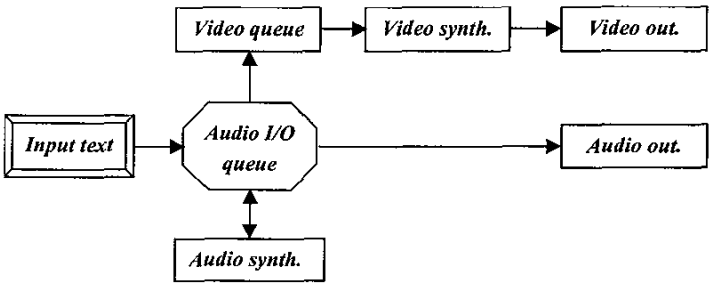

Figure 3: structure of the audiovisual synthesizer.

When the audio queue is full, the animation is started. The audio queue puts a couple of audio blocks to the audio device, which then starts to play the sound. The pre-synthesized image is swapped from the graphics card's memory at the appropriate time. Using the pre-calculated parameter set from the video queue, the next image is now synthesized. In addition to the speech parameters, also parameters for rotation, zoom, emotions and such are used for creating the new image.

The audio and video syntheses are always trying to keep their queues full. If the video synthesis takes more time than expected, the time to synthesize an audio block is increased and the audio queue starts to get smaller. When the size of the audio queue reaches a certain lower limit, we decrease the frame rate to give more processor time for the audio synthesis. In the opposite case, the frame rate is increased.

CONCLUSION

This new synchronization method gives us a better control over the audio-visual speech synthesis in time domain. Using, the diphone duration, we can use some more realistic interpolation function between the visemes. Then, we can take into account also coarticulation effects.

REFERENCES

[1] Sumby W. and Pollack I. "Visual Contribution to Speech Intelligibility in Noise", *Journal of Acoustical Society of America*, 26, 2, pp. 212-215, 1954.

[2] Beskow J., Dahlquist M., Granström B., Lundeberg M., Spens K-E and Öhman T. "The Teleface Project Multimodal Speech Communication for the Hearing Impaired" *Proceedings of Eurospeech '97*, pp. 2003-2006, Greece, 1997.

[3] Le Goff B., Guiard-Marigny T. and Benoît C. "Analysis-Synthesis and Intelligibility of a Talking Face" *Progress in speech synthesis*, J.P.H van Santen, *et al* (Eds), Springer-Verlag, pp. 235-246, 1996.

[4] Olivès JL, Möttönen R., Kulju J. and Sams M. "Audio-visual Speech Synthesis for Finnish" *AVSP'99, Proceedings of Audio-visual Speech Processing,* USA, 1999.

[5] Rubin P. and Vatikiostis-Bateson E. "Talking Heads" *Proceeding of Auditory-Visual Speech Processing,* Australia, 1998.

[6] Doenges P.K., Capin T.K., Lavagetto F., Ostermann J., Pandzic I.S. and Petajan E.D. "MPEG-4: Audio/video & Synthetic Graphics/Audio for Mixed Media" *Journal of Image Communications,* Vol. 5, No 4, pp. 433-463, 1997.

[7] Lukaszewicz K. and Karjalainen M." Microphonemic Method for Speech Synthesis" *Proceedings of IEEE ICASSP-87*, Dallas, USA, 1987.

[8] http://www.lce.hut.fi/researchface/demo.html

[9] Karjalainen M., Altosaar T. and Vainio M. "Speech Synthesis Using Warped Linear Prediction and Neural Networks" *Proceedings of IEEE ICASSP-98*, Seattle, USA, 1998.

[10] Strube H.W. "Linear Prediction on a Warped Frequency Scale" *Journal of Acoustic Society of America,* vol. 68, no. 4, pp. 1071-1076, 1980.

[11] Karjalainen M., Härmä A. and Laine U. "Realizable Warped IIR Filter and Their Properties'. *Proc. IEEE ICASSP-96*, Munich, Germany, 1996.

[12] Smith, J. O. and Abel, J. S. "The Bark Bilinear Transform. *Proceeding of IEEE ASSP Workshop*, USA, 1995.

[13] Parke F. "Parameterized models for Facial Animation", *IEEE Computer Graphics,* 2 (9), pp. 61-68, 1982.