DEVELOPING A 3D-AGENT FOR THE AUGUST DIALOGUE SYSTEM

Magnus Lundeberg and Jonas Beskow

Centre for Speech Technology, KTH, Sweden.
{magnusl, beskow}@speech.kth.se
http://www.speech.kth.se/multimodal/

ABSTRACT

In our continuing work with multimodal text-to-speech synthesis with high quality for speechreading, a new talking head has been developed with the purpose of acting as an interactive agent in a dialogue system, set up in a public exhibition area in downtown Stockholm. The new agent conforms to the same set of basic control parameters as our earlier faces, allowing us to control it using existing rules for visual speech synthesis. To add to the realism and believability of the dialogue system, the agent has been given a rich repertoire of extra-linguistic gestures and expressions, including emotional cues, turn-taking signals and prosodic cues such as punctuators and emphasizers. Studies of user reactions indicated that people have a positive attitude towards our new agent.

1. INTRODUCTION

Synthetic talking faces have been successfully employed in previous dialogue systems [1,2] at KTH and have also been used in different experimental studies on speech perception and human computer interaction [3]. The benefit in speech intelligibility of our previous lip-synchronized speech has been shown in the Teleface-project [4], both for hearing impaired persons and persons with normal hearing. Positive response from test subjects and other users have encouraged us to try to obtain a higher level of realism, suitable for new dialogue domains, as well as tools for hearing impaired persons. This paper describes how a synthetic face for lip-synchronized audiovisual output, based on the existing KTH audio-visual text-to-speech synthesis [5], was developed.

1.1 The August System

The August dialogue system [6] is a platform for presentation of the speech technology research carried out at the Centre for Speech Technology (CTT). One of the goals of the August project is to demonstrate how the speech technology modules developed at the lab can be put together to rapidly prototype a complex spoken dialogue system. The system was designed to handle a number of domains with different complexity rather than one single complex domain (such as e.g., ticket reservations). The simplest configuration of the August system is able to answer questions about Stockholm, the study program of KTH, the research at CTT or the about the system itself. August also has some knowledge about the author Strindberg and his literature.

The August system includes the following main components:

- A general purpose text-to-speech synthesizer
- A lip-synchronized 3D animated talking head described in this paper
- A general purpose speech recognizer for continuous speech
- Multiple dialogue managers, each dealing with it's own task domain
- An example based semantic analyzer with topic prediction
- A general broker architecture for handling the communication between the system's modules that are distributed over multiple platforms

The user utterances were processed by the speech recognizer, which generated N-best lists of the ten most probable utterances. These were analyzed to extract semantic information, such as domain, utterance type, and a set of semantic feature/value pairs. The topic prediction was used to determine which domain-specific dialogue manager to use. These dialogue managers were supposed to work independently to produce appropriate responses to send to the multi-modal synthesis module for generation. The main lexicon of the test-system included only about 500 words (of which several were multi-word expressions) and a bi-gram class grammar of 70 classes and 229 class pairs. For more details on the system and its components, see [6].

1.2 The quest for realism

The purposes of developing a new face were to make use of experiences from previous projects and to create a unique character for the August system. Opinions about our first generation synthetic face *Holger* from the hearing impaired subjects in the intelligibility tests of the Teleface-project were taken into account when developing the new face. Although many subjects expressed that they were more comfortable with the polygon-based synthetic face than with a photorealistic face, some improvement was wanted with regard to naturalness. The most frequent items on the wishlist were hair, ears and more realistic teeth.

For the August system we wanted an agent that looked like the 19th century Swedish author August Strindberg. The purpose of creating a Strindberg lookalike was

- To show a well-known character
- To indicate some knowledge about Stockholm, history and literature, thus implying the domain of the system
- To give the agent some personality. Strindberg is famous for making some rather categorical and well-known statements about politics, women, reviewers etc.

2. MODELLING AND SYNTHESIS OF TALKING FACES

Animated synthetic talking faces and characters have been developed using a variety of techniques and for a variety of purposes during the last two decades. For an overview of the field, see [7].

Our approach is based on parameterized deformable 3D facial models, controlled by rules within a text-to-speech framework [8]. The rules generate the parameter tracks for the face from a representation of the text, taking coarticulation into account [5].

We employ a generalized parameterization technique to adapt a static 3D-wireframe of a face for visual speech animation [9]. Based on concepts first introduced by Parke [10] we define a set of parameters that will deform the wireframe by applying weighted transformations to its vertices. One critical difference from Parke's system, however, is that we have de-coupled the model definitions from the animation engine, thereby greatly increasing the flexibility. For example, this allows us to define and edit the weighted transformations using a graphical modeling interface, rather than hand-coding them into e.g. C-source files. The parameterization information is then stored, together with the rest of the model, in a specially designed file format. For all our models developed to date, we have decided to conform to the same set of basic control parameters for the articulators that was used in [5]. This has the advantage of making the models independent of the rules that control them during visual speech synthesis; all models that conform to the parameter set will produce compatible results for any given set of parameter tracks.

The animation engine and the modeling interface currently run under Windows and several UNIX dialects including IRIX, Linux and HP-UX. Cross-platform portability was made possible by utilizing the industry standard OpenGL library for 3D-graphics and Tcl/Tk for the graphical user interface and scripting. For the August dialogue system we chose SGI/IRIX as the animation platform.

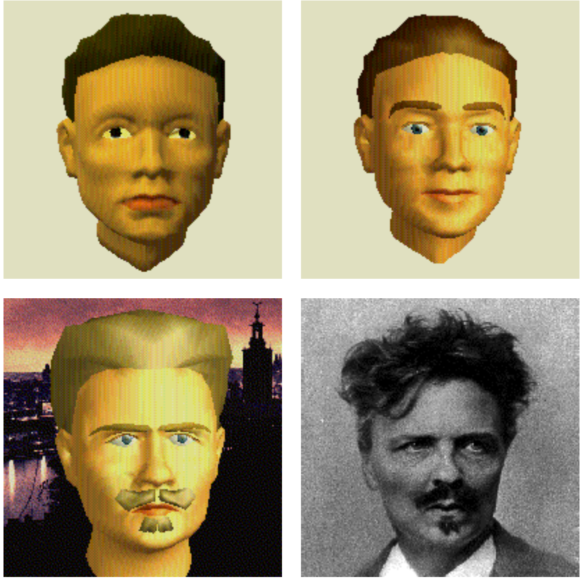

Figure 1: Top row: The static 3D-model from Viewpoint DataLabs, and the talking agent Alf. Bottom row: The talking agent August and the 19'Th century Swedish author August Strindberg.

3. DEVELOPMENT A NEW FACE

The new face was developed using a static 3D-model that was available as freeware from Viewpoint DataLabs [11] as a base. This 3D-model was built up by a reasonable number of polygons and head contours that could be made to look similar to Strindberg's with only a few adjustments (Figure 1).

Before parameterization could begin, the static viewpoint-model had to undergo some surgery. The lips were split to create a mouth opening, and the static eyes were replaced with moveable 3D eyeballs. We also added eyelashes, eyebrows, teeth and tongue, and increased the level of detail around the eyes. Teeth and the tongue were borrowed from the Holger model but enhanced to add to the realism. The teeth visible through the mouth opening were refined by adding polygons to give their surface a more convincing 3D appearance compared to the basically flat teeth used in *Holger* and *Olga*. The properties of the tongue were also refined to obtain a smoother look. It should be clearly stated that teeth and tongue were modeled to look good from the outside (i.e. through the mouth opening) and not to be anatomically correct when looking through the skin. An improved 3D-model of the internal articulatory organs [12] is currently under development at CTT. The face was parameterized using the methods described above and in [9]. The parameter set was chosen to conform to our earlier models, allowing us to control the face using existing target values for the visemes.

When applying the weighted transformations for the vertices, Vowel-Consonant-Vowel-words (VCV-words) from the Teleface project speech corpus [4] were used as prototypes. By viewing this speech material, both frame by frame and at full speed, and adjusting the weights in the model to resemble the articulation of our video recorded natural speaker, smooth and realistic articulatory movements were defined. This pre-August agent that was created for use as a general agent in different experimental setups was named *Alf*. After getting Alf up and talking with all the basic speech features the creation of *August* begun. The shape of the face was adjusted to make it more of a "Strindberg lookalike". A small beard and a mustache were added and August was also given some supernatural behavior like the ability to twist and stretch the beard and the mustache.

4. EMOTIONS, GESTURES AND IDLING MOTIONS

When designing the agent, it was considered of paramount importance that August should not only be able to generate convincing lip-synchronized speech, but also exhibit a rich and believable non-verbal behavior. To facilitate this we have developed a library of gestures that serve as building blocks in the dialogue generation. This library consists of communicative gestures of varying complexity and purpose, ranging from primitive punctuators such as blinks and nods to complex gestures tailored for particular sentences (Table 1). They are used to communicate such non-verbal information as emotion, attitude, turn taking, and to highlight prosodic information in the speech, such as stressed syllables and phrase boundaries. The parameters used to signal prosodic information in our model are primarily eyebrow and head motion.

Each gesture is defined in terms of a set of parameter tracks, which can be invoked at any point in time, either in-between or during an utterance. Several gestures can be executed in parallel. Articulatory movements created by the TTS will always supersede movements of the non-verbal gestures if there is a conflict. Scheduling and coordination of the gestures is controlled through a scripting language.

4.1 Prosodically motivated gestures

Having the agent accentuate the auditory speech with non-articulatory movements is found to be very important with respect to reactivity and believability of the system. The main rules of thumb for creating the prosodic gestures were to use a combination of head movements and eyebrow motion and maintain a high level of variation between different utterances. Earlier experiments [13], such as associating eyebrow movements with words in focal position, have indicated that this is a possible way of controlling the prosodic facial motion. Potential problems with such automatic methods are that the agent could look a little nervous or intense and that the eyebrow motion could become predictable.

To avoid predictability and to obtain a more natural flow, we have tried to create subtle and varying cues employing a combination of head and eyebrow motion. A typical utterance from August can consist of either a raising of the eyebrow early in the sentence followed by a small vertical nod on a focal word or stressed syllable, or a small initial raising of the head followed by eyebrow motion on selected stressed syllables. A small tilting of the head for-

Motion involved in gesture	**Typical usage**
Eyeblinks	Word boundaries, emphasis, idle random blinking.
Eye rotation	Thinking, searching
Head nodding	Emphasis, turn taking
Head turning	Spatial references (e.g. “the bathroom is over there”), attitude
Eyebrow raising	Mark words in focal position, stressed syllables, questions, emotions
Eyebrow frowning	Thinking, disagreeing
Twisting of mustache	Attitude e.g. proudness, aggressivness or flirting
Emotional expressions	Semanticaly motivated (See figure 2 for examples)

Table 1: Examples of typical usage for different motions when designing gestures for the dialogue system.

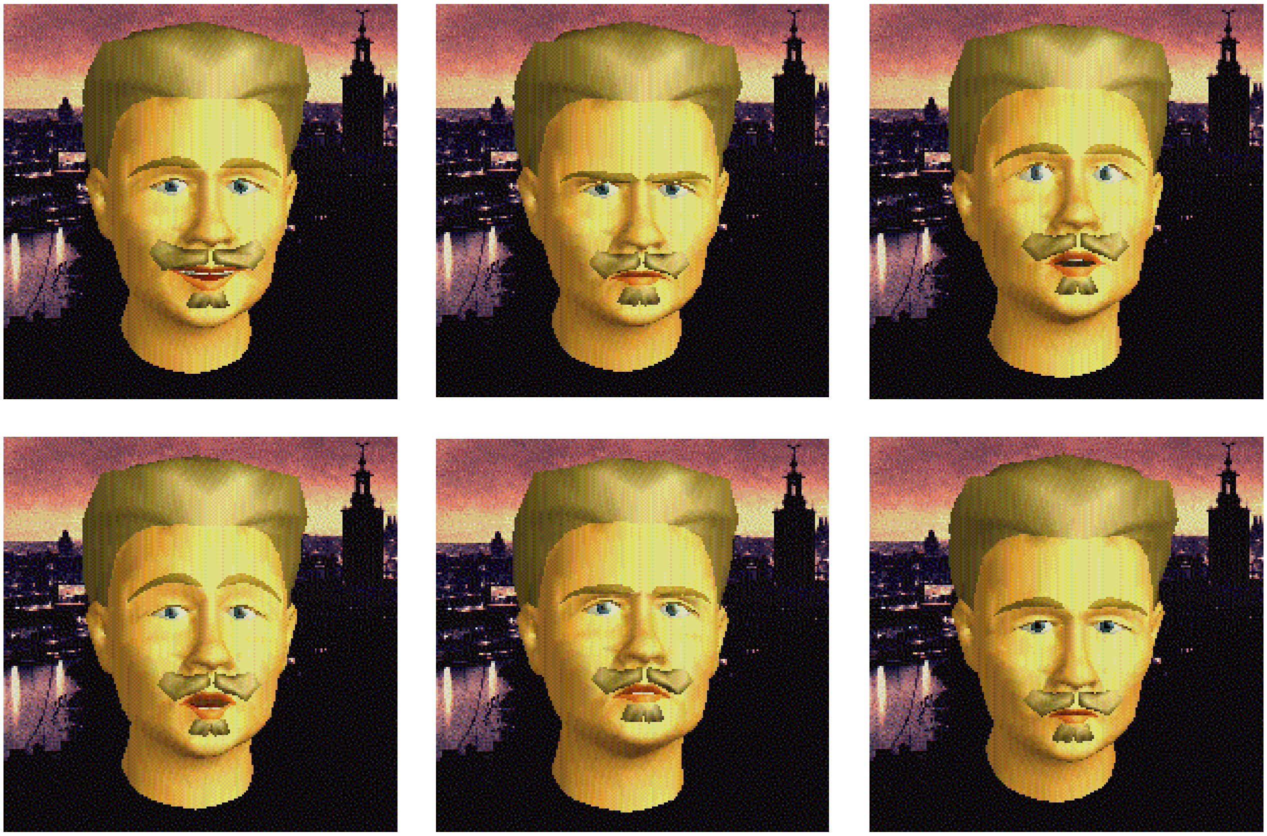

Figure 2: August showing different emotions: Happiness, Anger, Fear, Surprise, Disgust and Sadness.

ward or backward often highlights the ending of a phrase. A number of standard gestures with typically one or two eyebrow raises and some head motion were defined. The standard gestures work well with short answering sentences like “Yes, I believe so.” or “Stockholm is more than 700 years old.”

4.2 Agents have feelings too

To enable display of the agent's different moods, six basic emotions similar to the six universal emotions defined by Ekman [14] were implemented (Figure 2), in a way similar to that described by Pelachaud, Badler and Steedman [15]. Due to the limited resolution in our current 3D-model, some of the face properties, such as *'wrinkling of the nose'* are not possible, and were therefore left out. Appropriate emotional cues were assigned to a number of utterances in the system, often paired with other gestures.

4.3 An example sentence

Specially tailored, utterance specific gestures were created for about 200 sentences in the August system. An example of how eyebrow motion has been modeled in such a sentence is shown in Figure 3, and it can also be viewed on the CD-ROM from the workshop. The utterance is a Strindberg quotation [16] “Regelbundna konstverk bli lätt tråkiga liksom

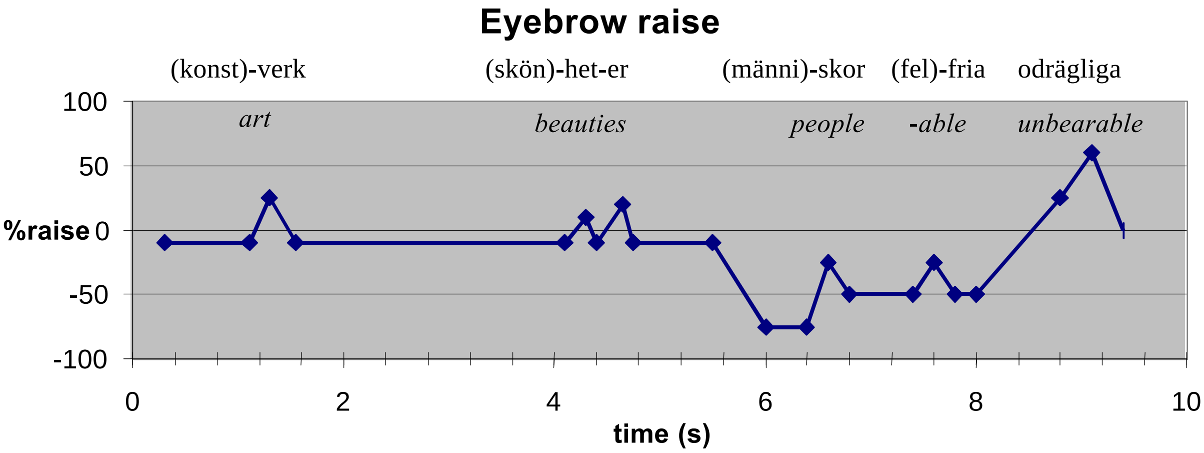

Figure 3: Example of eyebrow motion modeling in the August system. August says (in Swedish) “Regelbundna konstverk bli lätt tråkiga liksom regelbundna skönheter; fullkomliga människor eller felfria äro ofta odrägliga.”

regelbundna skönheter; fullkomliga människor eller felfria äro ofta odrägliga.” (Translation: *“Symmetrical works of art easily become dull just like symmetrical beauties; impeccable or flawless people are often unbearable.”*). Eyebrows are raised on the syllables *‘-verk’*, *‘-het’*, *‘-er’*, *‘-skor’* *‘-fria‘* and there is a final long rise that peaks on *‘a’* in the last word *‘odrägliga’*. Notice the lowering of the eyebrows that starts at t=5.8, which is not intended to convey prosodic information but rather emotional information. Not shown is the rotation of the head in the same utterance. At the first phrase boundary (after *‘tråkiga’* at t=3.1 s), August tilts his head forward. At the next phrase boundary (before *‘fullkomliga’* at t=5.5 s) he tilts the head even more forward and slightly sideways, lowers his eyebrows and looks very serious. A slow continuous raising of the head follows to the end of the utterance.

4.4 Listening, thinking and turn-taking

For turn-taking issues, visual cues such as raising of the eyebrows and tilting of the head slightly at the end of question phrases were created. Visual cues were also used to further emphasize the message (e.g. showing directions by turning the head). To enhance the perceived reactivity of the system, a set of listening gestures and thinking gestures was created. When the user presses the push-to-talk button, the agent immediately starts a randomly selected listening gesture, for example raising the eyebrows. At the release of the push-to-talk button, the agent changes to a randomly selected thinking gesture like frowning or looking upwards with the eyes searching. (Figure 4)

The August system was set up in an exhibition area at Kulturhuset (Stockholm Culture Center) in downtown Stockholm as part of the program of the *Cultural Capital of Europe '98*. To catch the attention of the people visiting the exhibition, various entertaining behaviors of the agent were needed. For example, August can roll his head in various directions, he can pretend he is watching a game of tennis and he can do a flirt gesture where he looks toward the entrance and performs a whistling while raising his eyebrows. These idling motions are important in making the system look alive, as opposed to a non moving, staring face which very well can make the system look 'crashed' or 'hung'. A high degree of variation for these idling motions is needed to prevent users from predicting the actions of the dialogue system.

5. CONCLUSIONS AND FUTURE WORK

The modeling of the gestures in the August project enhances the agent by making him act more natural and more convincing. While we don't claim that these motions are perfect, and certainly not the only possible ones, they do seem to provide the over all animation with some human-like qualities. Further work on how to generate the eyebrow and head motion will yield even more convincing interactive animated agents.

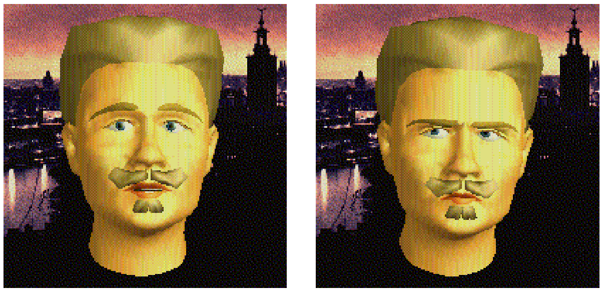
Figure 4: August listening (left) and thinking (right).

It is difficult to objectively judge the success of the agent in a system like this one – an open system with completely unsupervised user interaction. However, some indications might be given by studying the speech material collected in the August database [17]. It shows an incredible variety of questions that people have been asking the system. Although some of these questions obviously are stated to test the capabilities of the system, many of them are of high complexity and are stated by people expecting a good answer. This could indicate that the appearance and the audio-visual speech quality of our agent are believable enough to raise people’s expectations of the knowledge of the system.

The development of August and the August dialogue system has brought us knowledge of how to create 3D-agents as human-machine interfaces to dialogue engines with minor effort. A new agent named *Per* (Figure 5), was recently developed using the methods described here. Per is acting as a gatekeeper/receptionist in a current speaker verification/dialogue system project at CTT. A new version of Alf with a reduced number of polygons is under development, suitable for deployment on lower end

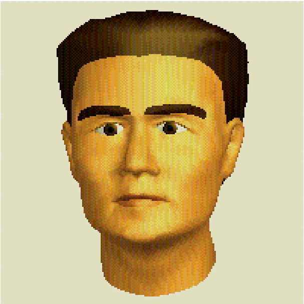
Figure 5: *Per* - the future gatekeeper/receptionist at CTT

platforms. Reduction is concentrated to areas with low surface curvature that were considered being of less communicative importance, such as the cheeks, forehead and the back of the head, while regions around the mouth and eyes are left intact.

Most of the gestures created for August can be used also with Alf or Per but the result will look less convincing. To increase the portability of gestures between our different models, the facial parameters defined in the models and the gesture procedures need to be more generalized. A gesture library is under construction, containing procedures with general emotion settings and non-speech specific gestures as well as some procedures with linguistic cues. These procedures can serve as a base for the creation of new communicative gestures in future dialogue systems.

Acknowledgments

The work with multimodal speech synthesis is done within the Centre for Speech Technology (CTT) at KTH. Some of the audio output of the text-to-speech synthesis used in the August system, is generated using an MBROLA database [18] for Swedish, created by Marcus Filipsson and Gösta Bruce of the Department of Linguistics and Phonetics of Lund University, Sweden.

References

1. Bertenstam, J., Beskow, J., Blomberg, M., Carlson, R., Elenius, K, Granström, B., Gustafson, J., Hunnicutt, S., Högberg, J., Lindell, R., Neovius, L., de Serpa-Leitao, A., Nord, L. and Ström, N. (1995): "The Waxholm system - a progress report", *In Proceedings of Workshop on Spoken Dialogue Systems*, Vigsø, Denmark, May 1995

2. Beskow, J., & McGlashan, S. (1997): "Olga - A Conversational Agent with Gestures", In *Proceedings of the IJCAI'97 workshop on Animated Interface Agents - Making them Intelligent*, Nagoya, Japan, August 1997.

3. Bengtsson, B., Burgoon, J.K., Cederberg, C., Bonito, J., Lundeberg, M. (1999). "The Impact of Antropomorphic Interfaces on Influence, Understanding, and Credibility". In *Proceedings of the 23nd Hawaii International Conference on System Sciences* - 1999, Maui, USA

4. Agelfors, E., Beskow, J., Dahlquist, M., Granström, B., Lundeberg, M., Spens, K-E & Öhman, T. (1998). "Synthetic faces as a lipreading support". In *Proceedings of ICSLP'98*, Sydney, Australia.

5. Beskow, J. (1995): "Rule-based Visual Speech Synthesis" In *Proceedings of Eurospeech'95*, Madrid, Spain, September 1995.

6. Gustafson, J., Lindberg, N. & Lundeberg, M. (1999) “The August spoken dialogue system” Accepted for publication in *Proceedings of Eurospeech’99*, Budapest, Hungary.

7. http://www.haskins.yale.edu/haskins/heads.html

8. Carlson R. and Granström B. 1997. Speech Synthesis. In Hardcastle W. and Laver J. (eds) The Handbook of Phonetic Sciences. 768-788. Oxford: Blackwell Publishers Ltd.

9. Beskow, J. (1997). "Animation of Talking Agents", In *Proceedings of AVSP'97, ESCA Workshop on Audio-Visual Speech Processing*, Rhodes, Greece, September 1997.

10. Parke, F.I. (1982) "Parameterized models for facial animation". *IEEE Computer Graphics. 2(9),* pp 61-68.

11. http://www.viewpoint.com/

12. Engwall, O. “Vocal tract modeling in 3D”. In TMH-QPSR 1/1999, Stockholm, Sweden.

13. Granström, B., House, D., Lundeberg, M. (1999) “Prosodic cues in multimodal speech perception” Accepted for publication in *Proceedings from ICPhS’99,* San Francisco, USA.

14. Ekman, P. (1979). About brows: Emotional and conversational signals. In M. von Cranach, K. Foppa, W. Lepinies & D. Ploog (Eds.), *Human ethology: Claims and limits of a new discipline: Contributions to the Colloquium* (pp. 169-248). Cambridge: Cambridge University Press.

15. Pelachaud, C., Badler, N.I. and Steedman, M. (1996). "Generating Facial Expressions for Speech" *Cognitive Science 28,*1-46.)

16. Strindberg, A. (1907). "A Blue Book"

17. Bell, L., Gustafson, J. (1999) “Repetition and its phonetic realizations: Investigating a Swedish database of spontaneous computer directed speech” Accepted for publication in *Proceedings from ICPhS’99,* San Francisco, USA.

18. T. Dutoit, V. Pagel, N. Pierret, F. Bataille, O. van der Vreken, "The MBROLA Project: Towards a Set of High-Quality Speech Synthesizers Free of Use for Non-Commercial Purposes" *Proc. ICSLP'96*, Philadelphia, vol. 3, pp. 1393-1396 .