DeMaMech 2005
Report

Freerk Wilbers

September 30, 2006

1 Personal Data

Name	Freerk Wilbers
Address	Brabantse Turfmarkt 20D
	2611 CN Delft, The Netherlands
Email	f.p.wilbers@student.tudelft.nl
Home Institute	Delft University of Technology
	Faculty of Mechanical, Maritime and Materials Engineering
	Department of BioMechanical Engineering
	Section Intelligent Mechanical Systems
	Mekelweg 2, 2628 CD Delft, The Netherlands
Supervisor	prof. dr. Tetsuo Tomiyama
Host Institute	Osaka University
	Graduate School of Engineering
	Department of Adaptive Machine Systems
	Intelligent Robotics Laboratory
Supervisor	prof. dr. Hiroshi Ishiguro

2 Executive Summary

This report describes my six-month exchange to Ishiguro laboratory at Osaka University, Japan. I left Holland in April 2006 and returned in October of the same year.

Ishiguro laboratory has designed an artificial human being, an android, a copy of a Japanese female newscaster named Repliee Q2. During my stay I worked on the generation of natural facial motions synchronised with and generated from speech.

In addition, living in Japan offered wonderful possibilities for exploring the country and its culture. I learned many things about Japan, Japanese people, and at the same time about myself during my stay here. I would heartily recommend a similar exchange experience to fellow students.

The report finishes with some concluding remarks, suggestions for improvements of the program, and some recommendations to future student. At the very end, a technical report detailing my work in Ishiguro laboratory is appended.

3 Travel Schedule

Amsterdam–Osaka	17 April 2006
Osaka–Amsterdam	14 October 2006

4 Technical Research Report

This section describes the research I carried out in Japan. The actual technical report is provided separately, in a paper format, at the end of this report. It includes some comments regarding future work. Please note that this report was handed in before the end of my stay, so my work was not quite finished at the time of writing.

4.1 Introduction to the Technical Report

The project I worked on was the generation of natural speaking motions in an android. The laboratory developed a number of androids, artificial human beings. They are pneumatically actuated, covered in silicone, have real hair and wear clothes. I worked with Repliee Q2, a copy of a Japanese newscaster with 42 degrees-of-freedom (DOF), of which 13 are in the face. The most recent android is Geminoid, a copy of professor Ishiguro himself at ATR, Advanced Telecommunications Research Institute, in Nara, Japan. Please refer to the laboratory website at `www.ed.ams.eng.osaka-u.ac.jp` for some pictures of Repliee Q2.

The android is to have sonzai-kan, or 'presence'. The challenge is to "quantify the elusive quality that makes people sit up and take notice, and figure out how it can be captured and transmitted." (from `wired.com`). From the research assignment description: "One of the main aims of humanoid robotics is to develop robots that are capable of interacting naturally with people. But to understand the essence of human interaction, it is crucial to investigate the contribution of behavior and appearance. Our group's research explores their relationship by developing androids that closely resemble human beings in both aspects."

In my research, I tried to achieve a part of this presence by trying to generate natural-looking facial motions synchronised with the speech. I analysed these motions using a motion capture system and tried to generate them in the android.

5 Exchange student life

This section describes some aspects of the life of an exchange student in Japan. Space is limited, so I will discuss only Japanese research culture, daily life in Japan, and the cultural and other leisure activities I undertook.

5.1 Research in Japan

Ishiguro Laboratory I stayed at professor Ishiguro's laboratory at Osaka University. Professor Ishiguro coordinates multiple research projects; the one I was involved in has been termed 'Android Science'. The aim is to build extremely life-like artificial humans (androids) such as *Star Trek*'s Data. About Android Science Tim Hornyak writes the following:

> To emulate human looks and behavior successfully, Ishiguro yokes robotics with cognitive science. In turn, cognitive science research can use the robot as a test bed to study human perception, communication and other faculties. This novel cross-fertilization is what Ishiguro describes as android science. In a 2005 paper, he and his collaborators explained it thus: "To make the android humanlike, we must investigate human activity from the standpoint of [cognitive science, behavioral science and neuroscience], and to evaluate human activity, we need to implement processes that support it in the android." [Hor06]

Perhaps because of his 'unconventional' research topics and approach professor Ishiguro is something of a celebrity in Japan – worldwide, actually – which was a lot of fun. He and his laboratory make frequent television (even Dutch national television) and magazine appearances. During my stay, I saw a number of film crews at work, including a Danish documentary film maker who stayed for several weeks.

In my research project I collaborated with dr. Carlos Ishi at ATR, the Advanced Telecommunications Research Insituture in Nara, where prof. Ishiguro has another research group. Carlos is an expert on natural speech processing. In addition to being able to benefit from his knowledge, the collaboration was a great chance to have a look around another prominent research institute.

Organisation The laboratory I stayed at, and the laboratories I visited, seem very well funded. There is an abundance of state-of-the-art equipment available to students.

Japanese universities – to the best of my knowledge – are not organised around sections or departments, but around laboratories. Laboratories are strongly linked to and often named after the professor that heads them. After their bachelor studies, students choose a professor they wish to study with. The severity of the obligatory entrance examination depends on the prominence and thus popularity of the laboratory.

Laboratory Students spend a lot of time in their laboratories (although that does not mean they are actually working). In my laboratory, 12-hour days and frequent weekend appearances were quite normal. A lot of this time is however spent chatting, organising activities, playing sports, reading comics, sleeping (frequently) and so on. The laboratory is above all a social place.

It is quite hierarchically and formally organised. First, functioning as a group is very important. Trips, drinking parties, welcome parties and farewell parties are held (often they are a little more 'formal' than Western visitors are used to) to promote the lab 'structure'. An example of this formality is that Japanese parties will have an ending time as well as a strict starting time. One starts together, drinks together ('kampai!'), and cleans up together.

Second, there is something of a hierarchy in the lab. Staff 'outranks' students, doctoral students 'outrank' master students, who 'outrank' bachelor students, and so on. This is reflected in the way students approach (or do not approach) each other. Foreign exchange students do not really 'fit' in this hierarchy, which is sometimes confusing for the Japanese students, but also allows for more informal contacts.

Robotics in Japan Japanese robotics research is ambitious and long-term oriented. Many of the research projects seem outlandish, odd, or even worrying to foreigners – take for instance Paro, the robotic baby seal designed as a therapeutic companion for socially isolated individuals.

Japanese seem to have a very different and much more positive attitude to robots, and technology in general – robotics is considered essential in battling Japan's social problems, for instance supporting the rapidly aging population. For an excellent analysis of the above, see [Rep05]. These factors make Japan an extremely fascinating place for research.

5.2 Daily life in Japan

Daily life in Japan is surprisingly easy to cope with, even without speaking the language. Supermarkets and all-night convenience stores are everywhere. Trains are easy to use, quite affordable (although the *shinkansen* bullet trains are terribly expensive) and extremely punctual. Beware that they stop running shortly after midnight.

It helps to speak a little Japanese. People will appreciate it – even the tiniest amount of spoken Japanese will earn visitors tremendous praise – and also, the amount of English spoken decreases rapidly as one leaves Tokyo. My Japanese never progressed much beyond greetings, names of foods, restaurant-related phrases and occasional bits of slang. I must admit I did not take any Japanese language lessons, as the level of the 'casual speaking' courses offered by the university was very low, and the intensive courses are very time-consuming. Students and researchers intending to work or study in Japan usually take a six-month intensive (full-time) Japanese course, after which they can communicate in basic spoken and a little written Japanese. Obviously this is not feasible if one is staying only six months.

Japan has a reputation for being expensive. In fact, although rent and travel can be very expensive, daily commodities and especially foods are quite cheap. It is perfectly possible to survive on a normal Dutch university student's budget, although for anything beyond work-dormitory commutes and simple meals, such as sightseeing, entertainment, gifts, clothes, the extra budget provided by the exchange program was necessary.

In short, daily life in Japan is so 'normal' that it soon stops feeling like a foreign country – although exactly this normality can start to feel a little unsettling.

Dormitory I'd like to spend a little extra time discussing the dormitory DeMaMech students were assigned in Osaka. The Senri International House of Osaka Prefecture is conveniently located two stations from the university campus, has relatively large rooms, and by Japanese standards the rent is cheap. On the other hand though, its condition, its rules and its atmosphere make it a place that, in my opinion, is unsuitable for living. The standard of living in Senri International House is far below what one expects of a modern, civilised country and far below Japanese standards of living. I will briefly discuss some features of the dormitory.

The building has the oppressive look, feel and smell of a seventies prison hospital. It is cold and windy in winter, hot and stuffy in summer. Shared toilets, kitchen and showers are incredibly dirty and badly equipped. Cleaning of facilities and hallways consists of dragging a filthy wet rag over some of the visible surfaces. Cockroaches are everywhere, including residents's rooms. Four coin-operated showers on the ground floor are shared between 56 men living across five floors.

Despite working in an international house, the staff does not speak English. Communication with them is limited to notes written in Japanese. Visitors (including other residents of the dormitory) are not allowed in residents's rooms. A one-o'clock curfew interferes with social and research schedules. An eleven-o'clock lockup of the washing machine facility, the recreation room and the study room confines students to their rooms and makes the simple act of doing the laundry while having a busy schedule an exercise in logistics. Many scholarship students in the dormitory live in permanent fear of breaking one of the many rules, being reported by the ever-vigilant guards to their institutions and having their scholarships revoked.

One can understand that these conditions do not make the dormitory a cheerful place to return home to after a long day in the laboratory. In fact, I don't know anyone who called it home – it was referred to as 'the dormitory', at best. Osaka University and the city of Osaka offer far better and equally affordable dormitories and similar housing facilities for international students – for bureaucratic reasons it was impossible to move in to one of them. I strongly suggest that future DeMaMech students – in fact, all exchange students – be housed in one of these other facilities.

5.3 Culture and leisure

There is really no space to describe the many fascinating things I learned about Japanese culture. After one has experienced the normality of daily life mentioned earlier, the little discoveries of hidden differences take over until once again, Japan becomes a novel experience. Discussing these discoveries with other foreigners is a very infectious habit. As a result, foreign writers have devoted countless pages to describing and analysing Japan; I'll refrain from doing it all over again.

As for leisure activities, I took many trips around the country. I visited Tokyo and nearby fabulous Nikko together with a visiting Dutch friend, age-old Kyoto, ancient Nara, modern Hiroshima and beautiful Miya-jima with my parents, tropical Okinawa with fellow DeMaMech students ,and many other places, alone and with others. Osaka itself is a vibrant and bustling city in which there is more than enough to explore without ever needing to set foot in Tokyo. Japanese onsen, hot springs, are a wonderful experience, as well as the more day-to-day sento baths, of which I have become a great fan.

Japanese food is tasty and healthy, and much more varied than the image of sushi and teriyaki that persists in the West. Japanese local beer is excellent – beware however of the cheaper, tax-dodging and headache-inducing variations on beer ('happoshu') made from beans, rice, or anything in between. An izakaya, a restaurant-bar blend that comes in many varieties is a great way to spend a night out with friends. After the izakaya, a visit to the karaoke bar is a classic. Your group rents a private room and the karaoke machine corrects your voice, so the risk of embarrassment is limited.

I did spend most of my leisure time with fellow foreign students as it is difficult to befriend Japanese (language, culture, and other barriers). One exception is the 'tutor' assigned to me by the lab, Kayo Yoneda, who took me on many wonderful trips and introduced me to her family and friends.

6 Concluding remarks

Studying in Japan is a unique experience. Studying overseas is rewarding in its own right; studying in Japan is a little more rewarding. I learned a lot about its history, its culture and learned to understand its people. Robotics research in Japan is advanced, occasionally bewildering, sometimes even a little crazy, but always fascinating.

Working in Japan is very different from working in Europe. There is a language barrier, which means that the experience and knowledge your labmates have may be difficult to access. Also, Japanese will generally be less critical of your work, and hesitant to offer their advice. I discovered only later on in my stay that a greater degree of independence and assertiveness is required to obtain the information and guidance one requires.

A second difficulty I experienced is with planning. When starting a new project, it is easy to underestimate the amount of work that needs to be done, resulting in plans that are plainly unfeasible. Six months is a short time in which to start a new research topic; I advise future students to take this into account better than I did.

The third difficulty, or rather disappointment, is one already mentioned and concerns our accommodation. My stay would have been that much more pleasant had I been able to live in a proper apartment or reasonable dormitory – I feel that in order to work successfully in a new country, one needs a place one can call home.

Despite these occasional difficulties, I enjoyed my stay in Japan tremendously. The rewards of this exchange far outweigh the troubles on the way.

I would like to thank DeMaMech's supervisors, especially prof. Tomiyama and prof. Fujita for offering this opportunity. Special thanks go to my supervisors prof. Ishiguro, my Delft supervisor, who happens to be prof. Tomiyama, and ass. prof. Minato. Also I would like to thank the members of Ishiguro laboratory, especially my tutor and friend Kayo Yoneda. Finally I would like to thank in particular Andrej, Hareld, Maarten, Zen and Yin for the wonderful time we spent in Japan together.

References

[Hor06] Tim Hornyak. Android science. *Scientific American*, May 2006.

[Rep05] Special Report. Better than people. *The Economist*, December 20th 2005.

Human-like speech in an android

Freerk Wilbers
f.p.wilbers@student.tudelft.nl

Abstract

We present an outline of a method for creating human-like facial motions driven from speech in a life-like humanoid robot, an android. We propose to separate this complex problem in to two subproblems, one being prediction of motion parameters speech, the other being generation of natural motions in the android. We present and discuss results of motion transfer from human to android by two different methods. Finally, future work is outlined.

1 Introduction

Ishiguro laboratory, in collaboration with Kokoro, Ltd., has developed an artificial human being, an android named Repliee Q2. Repliee Q2 is a 42-degree-of-freedom pneumatically actuated model of a Japanese female, Fig 1. For this android to be accepted by humans as a natural conversation partner, it must exhibit behaviour similar to that of humans.

In this work we aim to develop a part of these natural speaking motions; the part that is associated with the production of speech. It has been shown, for instance [1], that there exists a strong correlation between features of the speech signal and corresponding movements of the face and vocal tract.

The source of this signal can be either a text-to-speech system (speech synthesis) or from a human being. In this manner we can create an expressive virtual presence (telepresence).

A similar application, which has been researched extensively, is the creation of so-called ‘talking heads’ – virtual agents or animation characters with lip-synchronised speech. Many researchers have worked on predicting facial expression parameters from speech, using Hidden Markov Models, Neural Networks, and other methods , for instance [2, 3, 4, 5, 6, 7]. These parameters can be used to drive animations, but also to drive more sophisticated mechanical models of the human face, such as developed by Waters [8], for instance.

However, in all of these cases the driven platform is virtual, meaning that any desired deformation of the face can be generated. In our case, however, we are dealing with a physical platform, with a limited number of actuators, a limited range of motion and limited velocities, a silicone skin that does not behave like human skin, and so on. The question is how we can generate human-like motion on a platform with these limitations.

There exist some other physically-embodied ‘talking heads’; for example Kismet by Cynthia Breazeal [9] and its successor Leonardo. The degree of synchrony and expressiveness in these cases is however limited, and the motion generation method used is quite basic. Kismet opens its mouth as voice power increases, and closes the mouth for plosive (e.g. /p/) sounds.

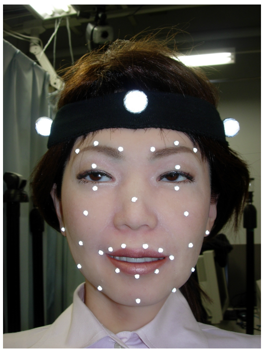

Figure 1: Markers applied to the android

Examples exist of the transfer of human motion to robot motion using inverse kinematics models. Yamane et al. [10] transferred human motion to a simple marionette, Pollard et al. [11] to a humanoid robot. In all cases, however, the motions were ‘large-scale’ motions of limbs and torso, not the subtle and complex motions in the human face.

The present work focuses on transferring those motions that are associated with the production of speech. In future works, we would like to investigate the additional transfer of emotional and other facial expressions.

2 Approach

2.1 Data Collection

In experiments, data from two human subjects was captured, in addition to facial motion data of the android. Motion data was captured by applying 31 small reflective markers on the face and tracking them with a motion capture system.

For the human subjects, segments of conversation (with another human participant) were recorded, as well as the subject reading out loud some prescribed texts. For each subject, approximately half an hour of data was collected. For the android, the contributions of the actuators to facial deformations were recorded.

In addition to motion capture data, video and audio files were recorded for the human subjects. The marker layout is given in Fig. 1.

2.2 Problem decomposition

We decompose the problem to form two separate, but related problems. The first is finding a mapping between the human speech signal, for instance encoded as a wave file, and the corresponding facial motion. Some issues in this problem are how to parameterise speech, whether or not to include language ‘understanding’, whether or not to use a

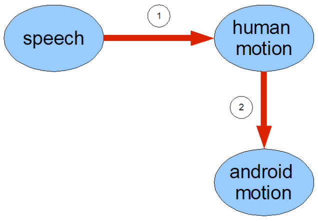

Figure 2: Schematic approach

physical model of speech generation, and so on. As described in the introduction, this is not a ‘new’ problem.

The second problem is finding a mapping between the human facial motion and the android facial motion. The issues we deal with here are different: for instance the different structure of the android face, the actuator dynamics and so on. It therefore makes sense to separate these problems. Fig. 2 is a schematic representation of the approach.

3 Speech to human motion

There are basically two approaches to facial motion generation [12]: one is language-based (extraction of phonemes; generally considered basic units of speech), e.g. [2] and the other is more physics-based (extraction of low-level parameters), e.g. [1, 13].

Phoneme extraction, although popular, suffers from the problems of natural speech processing: it is difficult and not robust. In addition, information may be lost in classification of phonemes. Therefore, we choose the second method.

Co-articulation Human speech is not neatly divided into easily discernible units. Sounds, and mouth shapes, depend on preceding and subsequent sounds and shapes. For example: the shape corresponding to ‘f’ in ‘far’ is different from the shape for the same letter in ‘for’, because of the vowel that follows it. This effect is termed *co-articulation* and it occurs on different levels and different timescales [3]. A serious lip-synching attempt should take this effect into account. The first, phoneme-based, approach described above does this by looking at two- or three-phoneme sequences (triphones). The second, physics based, approach considers not only the current frame, but also its neighbouring frames.

3.1 Implementation attempts

Speech recorded from the human subjects is parameterised with LPC Cepstral coefficients, a frequency-based approach that is common in speech processing. From these coefficients, we attempt to predict corresponding human facial motions (parameterised as the movement in 3D-space of the 31 facial markers).

Before attempting prediction, it is useful to reduce the dimensionality of the human facial motion, currently 91. A common technique is Principal Component Analysis, which has been used by other researchers to successfully reduce the dimensionality of human motion data, e.g. [6]. We found that the first 7–10 principal components of facial motion can represent 90–95% of variance.

A first prediction attempt was made using a linear regression model with 7 terms (to account for the effect of co-articulation introduced earlier), Fig 3. It shows that although we can roughly predict lip height (indicating the correlation shown in [1]) we cannot predict it anywhere near accurately using a linear model.

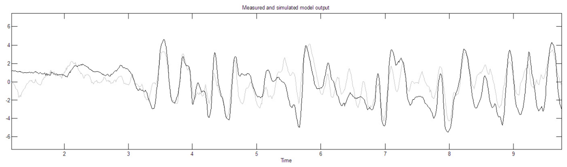

Figure 3: Performance of linear model in predicting human mouth height from speech, on validation data

Currently we are investigating neural networks with a speech data window as input (again to account for co-articulation). So far, they are not able to sufficiently generalise for new speech, even with large hidden layers. Researchers have successfully used a combination of methods, for instance first classifying speech into phonemes and then processing each phoneme class with its own neural network [6].

More work is obviously needed on this problem.

4 Human motion to android motion

Basically, we are trying to copy human facial motion to the android. Human and android motion are represented as the three-dimensional position in time of 31 markers – a 91-dimensional space.

One of the issues with copying motion is the different facial geometries of human and android. Another is how to represent human motion: because of the different geometries, absolute marker coordinates are difficult to deal with.

Because of these different facial geometries (and no two humans have the same face) we suppose that these absolute coordinates are not that important. Rather, we must look at timing, velocities and accelerations, motions relative to facial features and so on.

In addition, the android has some physical limitations. Among these are a limited number of degrees-of-freedom (DOF), less-than-ideal actuator behaviour (dead time, hysteresis), a very different facial tissue structure and no possibility of real-time control. A mapping must take into account these limitations.

The android’s facial actuators are shown schematically in Fig. 4. There are 13 facial actuators, of which two, numbers 3 and 4, move the eyes. As these cannot be motion-captured using reflective markers, we disregard them in this work. This leaves 11 facial actuators to be controlled. Actuator number 1 raises the eyebrows, number 2 blinks, number 5 opens the eyes wide expressing surprise, no. 6 moves the jaw, nos. 7–10 move the upper and lower lips, nos. 11 and 12 dimple the cheeks, and no. 13 raises the cheeks in an expression that resembles anger.

Normalising data As mentioned, we are considering only motions in the face associated with speech production. Head and shoulder motions by the subject during the motion capture sessions (section 2.1) introduce undesired scaling and rotation. To remove these we use a scaling and rotating approach based on singular value decomposition described in [14].

In the next sections we will describe two experiments. Video’s of the results described are available online at

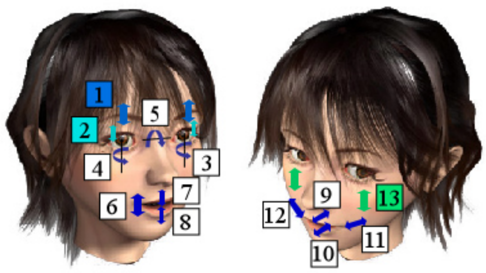

Figure 4: Android facial actuators

www.ed.ams.eng.osaka-u.ac.jp/amsed049/.

4.1 Position mapping by error minimisation

In this experiment we consider only the nine markers around the mouth. We copy a mouth shape to the android in a motion capture experiment by minimising the error between the android and target marker positions. The minimisation algorithm used is a hillclimbing algorithm developed by Takashi Minato.

The error to be minimised is given by

$$\epsilon = \sum_{i=1}^{3n} \|x_i^{tar} - x_i^{real}\|^2 + n \cdot \mathrm{std}\|x_i^{tar} - x_i^{real}\|, \qquad (1)$$

where x_i is a coordinate in space and n is the number of markers. In this manner we copy about 100 mouth positions.

We can generate a motion sequence by selecting significant mouth positions (‘keyframes’) and stringing them together by interpolating between them.

4.1.1 Evaluation

The resulting motion looks a little ‘odd’. For instance, the upper lip moves in a very jerky, discontinuous fashion. The motion is not smooth. Also, as the solutions were constrained only by the mouth markers, the algorithm was allowed to move around the jaw at will. It used the jaw actuator as its primary control, and neglected the lips. Also, the hillclimbing part of the method is computationally quite expensive.

4.2 Kinematic model

To try and solve some of the problems described above, a simple kinematic model of the android is constructed. It should provide some higher-level knowledge of actuator behaviour. Face configurations are described by a 91-dimensional vector $\mathbf{x}$, the concatenated 3D coordinates of the 31 markers. Android actuator settings are described by 11 activation parameters α_i, $0 \leqslant \alpha_i \leqslant 1$.

Face transform The geometries of the human and android faces are different, so there will not be a one-to-one correspondence between the two marker sets. We use a static transform defined as:

$$\mathbf{x}_{transform} = \hat{\mathbf{x}}_{android} - \hat{\mathbf{x}}_{human}, \qquad (2)$$

where $\hat{\mathbf{x}}_{android}$ and $\hat{\mathbf{x}}_{human}$ are hand-picked neutral positions of the android and human faces, respectively.

We motion-capture the android face with all its actuators in the rest position. We then move actuators one at a time and capture the deformed face when the actuators are at their maximum positions. We do this for the 11 actuators used in this work.

From these deformed faces, we subtract the neutral android face $\hat{\mathbf{x}}_{android}$ to obtain for actuator i its contribution $\mathbf{x}_i$. Assuming that the relation between displacements and actuator activation is linear, and that all actuators are independent, we write

$$\mathbf{x}_{predict} = \left[\sum_{i=1}^{\text{no. DOF}} \alpha_i \cdot \mathbf{x}_i\right] + \hat{\mathbf{x}}_{android}, \qquad (3)$$

where α_i is the activation of actuator i, $0 \leqslant \alpha_i \leqslant 1$. This equation models how the android face is deformed for a given combination of actuator settings α.

The error, then, between the actual human face and that predicted in the android is

$$\epsilon = \mathbf{x}_{actual} + \mathbf{x}_{transform} - \mathbf{x}_{predict}, \qquad (4)$$

where $\mathbf{x}_{transform}$ is the static transformation between human and android face of Eq. (2). The combination of actuator settings α that approximates the human face best is found by minimising the error given above. For this we use a bounded (as α must lie between 0 and 1) linear least-squares routine in MATLAB.

Doing this for every frame in the sequence gives a playable sequence.

4.2.1 Evaluation

The predicted actuator activations are generally continuous in time, even though they are estimated frame-by-frame. This indicates the validity of the method. The resulting motion is smoother and produces more realistic facial motion than in the previous experiment. The mapped expressions match those in the human subject, within the constraints imposed by the android’s mechanical construction. Not all of the human subjects fast motions can be transferred, due to the limited speed of the android actuators, but the overall similarity of the motion is convincing.

5 Conclusion

We have not yet been able to generate facial motions directly from speech. However, we have successfully and convincingly mapped human facial motion sequences to the android. These are promising steps in the creation of natural facial motions in the android. The Japanese speaking motions in the android are lively and both statically and dynamically very similar to human motions. In an informal evaluation, several Japanese speakers agreed that the produced motion looked ‘correct’ in its timing and general facial shapes.

Of course, a great deal of work remains to be done, some of which is described in the next section.

6 Future work

In the next few weeks, we will attempt to identify some important factors in the human-likeness of the generated motions by comparing sequences generated in various ways. Our model so far has been a linear model. We would like to try and estimate the remaining non-linearity of the human facial motion. Below we will outline some other issues that remain.

6.1 Regarding motion transfer

Optimisation of the generated sequence A Simulated Annealing algorithm was developed that attempts to optimise the generated sequence by varying some of its parameters. The system plays back a sequence, motion-captures it, and evaluates it by comparing the generated marker paths with the reference ones. We would like to use this algorithm to 'tweak' the sequence, but the question is which parameters to modify – timing, magnitudes or others.

Facial transform The facial transformation, Eq. (2), is a static one based on only the neutral faces. However, not only the geometry, but also the kinematics of the android face are different from the human face. A start for a more realistic transform could be to 'match' not only the neutral face, but other key expressions (such as the basic human facial expressions) to hand-selected ones in the android face, to create not a static mapping, but an interpolated 'mapping space'.

More advanced inverse kinematics A more realistic inverse kinematics model can be obtained by estimating the jacobian relating motion of end points (the markers) and actuator activation,

$$\Delta \mathrm{M} = \mathbf{J} \cdot \Delta \mathrm{x}.$$

For inverse kinematics and motion transfer to robots see for instance [10].

6.2 General

Obviously, the mapping from speech needs to be completed. Some other general issues that remain are below.

Assessment of motion quality In a more formal evaluation of the quality of the generated motion, we would like to investigate whether the produced motions contribute test subject's perception of the android's lifelikeness. The laboratory has previously carried out such experiments to assess other aspects of the android's behaviour.

Expression of emotions and other motions As mentioned in the introduction, we would like to extend the mapping to the expression of motions and the incorporation of other 'subconscious' gestures – nodding, for instance. For instance head motion has been emperically shown to be correlated to the pitch of the speech signal [15].

Mapping speech directly to actuators We mentioned previously that PCA was applied to the 91-D marker coordinate space to reduce its dimensionality. Mapping motion to the android is in itself a reduction of dimensionality, as we are mapping into the limited actuator space. We would like, therefore, to map directly from the audio signal into the android actuator commands.

References

[1] H. Yehia, P. Rubin, and E. Vatikiotis-Bateson, "Quantitative association of vocal-tract and facial behavior," *Speech Commun.*, vol. 26, no. 1-2, pp. 23–43, 1998.

[2] T. Ezzat, G. Geiger, and T. Poggio, "Trainable videorealistic speech animation," in *Proceedings of SIGGRAPH 2002*, (San Antonio, TX), 2002.

[3] M. Brand, "Voice puppetry," in *Proc. SIGGRAPH '99*, pp. 21–28, 1999.

[4] D. W. Massaro, J. Beskow, M. M. Cohen, C. L. Fry, and T. Rodriguez, "Picture my voice: Audio to visual speech synthesis using artificial neural networks," in *Proc. AVSP'99* (D. Massaro, ed.), (Santa Cruz, CA), pp. 133–138, August 1999.

[5] F. Lavagetto, "Coverting speech into lip movements: A multimedia telephone for hard of hearing people," *IEEE Transactons on Rehabilitation Engineering*, vol. 3, pp. 90–102, March 1995.

[6] P. Hong, Z. Wen, and T. S. Huang, "Real-time speech driven face animation with expressions using neural networks," *IEEE Transactions on Neural Networks*, vol. 13, pp. 916–927, July 2002.

[7] T. Kuratate, H. Yehia, and E. Vatikoitis-Bateson, "Kinematics-based synthesis of realistic talking faces," in *Proc. Auditory-Visual Speech Processing, AVSP98*, 1998.

[8] K. Waters, "A muscle model for animation three-dimensional facial expression," in *SIGGRAPH '87*, (New York, NY, USA), pp. 17–24, ACM Press, 1987.

[9] C. Breazeal, "Emotive qualities in lip-synchronized robot speech," *Advanced Robotics*, vol. 17, pp. 97–113, May 2003.

[10] K. Yamane, J. K. Hodgins, and H. B. Brown, "Controlling a marionette with human motion capture data," in *Proc. ICRA03*, pp. 3834–3841, 2003.

[11] N. S. Pollard, J. K. Hodgins, M. J. Riley, and C. G. Atkeson, "Adapting human motion for the control of a humanoid robot," in *Proc. ICRA02*, vol. 2, pp. 1390–1397, 2002.

[12] T. Chen and R. R. Rao, "Audio-visual integration in multimodal communication," *Proceedings of the IEEE*, vol. 86, pp. 837–846, May 1998.

[13] R. Gutierrez-Osuna, P. Kakumanu, A. Esposito, O. Garcia, A. Bojorquez, J. Castillo, and I. Rudomin, "Speech-driven facial animation with realistic dynamics," *IEEE Transactions on Multimedia*, vol. 7, pp. 33–42, February 2005.

[14] M. B. Stegmann and D. D. Gomez, "A brief introduction to statistical shape analysis," *published online*, 2002.

[15] H. Yehia, P. Rubin, and E. Vatikiotis-Bateson, "Facial animation and head motion driven by speech acoustics," in *5th Seminar on Speech Production: Models and Data* (P. Hoole, ed.), Ed. Kloster Seeon, Germany, 2000.