ACTA ACUSTICA UNITED WITH ACUSTICA
Vol. 90 (2004) 1084 – 1095

From Audio-Only to Audio and Video Text-to-Speech

Eric Cosatto*, Hans Peter Graf*, Jörn Ostermann†
{eric.cosatto, hpgraf, joern.ostermann}@ieee.org

Juergen Schroeter
AT&T Labs – Research Room D163, 180 Park Ave, Florham Park, NJ 07932, USA. jsh@research.att.com

Summary
Assessing the quality of Text-to-Speech (TTS) systems is a complex problem due to the many modules involved that address different subtasks during synthesis. Adding face synthesis – the animation of a "talking head" and its rendering to video – to a TTS system makes evaluation even more difficult. In the case of talking heads, today, we are at the infancy of research towards evaluating such systems. This paper reports on progress made with the AT&T sample-based Visual TTS (VTTS) system. Our system incorporates unit-selection synthesis (now well known from Audio TTS) and a moderate-size recorded database of video segments that are modified and concatenated to render the desired output. Given the high quality the system achieves, we feel for the first time that we are close to passing the Turing test, that is, that we are almost able to synthesize "talking heads" that look like recordings of real people. We demonstrate this point in applications, either over the web (client/server), or in stand-alone form, in a kiosk setting. Several steps are necessary to assure a very high quality sample based VTTS system. First, highly accurate image analysis tools are important for creating the necessary video clip databases. The problem is compounded by the fact that facial videos cannot be stored whole due to unfavorable combinatorics: for a given synthetic sequence, it is very unlikely that a whole face video clip contains the correct mouth sequence, the appropriate eye sequence, and also a suitable "background" face, given what we want to synthesize. Consequently, separate parts of a synthetic face need to be accessible independently from each other at synthesis time. Therefore, image analysis tools semi-automatically extract (i.e., cut) desired facial features out of recorded video, normalize the apparent position of the camera (the "pose", i.e. angle and distance between face and lens), and index and store the images in disjoint databases. Second, fast search techniques ("unit selection") extract the most appropriate sequences of facial building blocks at runtime. This includes background face images that convey desired head movements and serve as canvases for painting (projecting) other content-bearing parts of the face such as mouth and eyes. In a final step, the resulting composite face image is then rendered on a graphic screen for display. The higher the quality of a (V)TTS system, the more important it is to carefully evaluate all algorithmic choices. Naturally, subjective testing, although time consuming and expensive, has to be the ultimate measure. However, we used objective measures for quality assessment during the development phase of our system. For example, we found that accuracy and timeliness of lip closures and protrusions, turning points (where a speaker's mouth changes direction from opening to closing), and overall smoothness of the articulation are very critical for achieving high quality. We also found that "visual prosody", the movement of the head in synchrony with the stress pattern of the spoken sentence, is important for a natural look.

PACS no. 4372.Kb, 43.72.Ja, 43.71.Gv

1. Introduction

At the start of the new millennium, telecommunications has fully embraced Internet-Protocol (IP) networks in form of supporting multiple media such as voice, video, documents, database accesses, etc. Going forward, more and more devices, from telephones to PDAs and PCs, will enable communications over IP networks in multiple modalities, including "video" in addition to the traditional "voice" communication. Increasingly, human-to-human communication will be amended by communication between humans and machines for such applications as e- commerce, customer care, and information delivery in services [1].

The "speech circle" depicted in Figure 1 illustrates the general concepts and different modules used in natural lan-

Received 8 November 2003,
accepted 14 April 2004.

* Currently at NEC Laboratories, Princeton, New Jersey

† Now with Institut für Theoretische Nachrichtentechnik und Informationsverarbeitung, University of Hannover, Germany

guage voice interactions with machines. The process can be summarized as follows. The speech signal related to a customer's (top center) voice request is analyzed by the Automatic Speech Recognition (ASR) subsystem shown on the top right. The decoded words are input into the Spoken Language Understanding (SLU) component. The task of the SLU component is to extract the meaning of the words. For example, the recognized words: "What discounts are available?" imply that the customer would like to find ways to lower the price for a product or service. Next, a Dialog Manager determines the next action the speech-enabled customer-care system should take ("determine whether item is on sale or educational and other discounts are available.") Let's assume that a database lookup reveals that only an educational discount is available. Consequently, the SLU instructs the Language Generator (LG) to construct a reply (words, and instructions on how to say the words). Finally, the Text-to-Speech (TTS) component synthesizes the question "Are you a member of a school, or other educational institution?" – starting another turn of a multi-turn dialog. Note that we may categorize the ASR and TTS modules as "speech engines", while the SLU, DM, and LG modules encapsulate the "artificial intelligence" of the system.

The quality of the synthesized speech output is very important. Literally, "TTS is closest to a customer's ear". Fortunately, TTS systems with previously unavailable naturalness are becoming ubiquitous.

Introducing the topic of this paper, Figure 1 also illustrates a possible extension to the voice-only speech circle: high-quality "talking heads" – the visual rendering of animated (synthesized) head-and-shoulder images, also called Visual Text-to-Speech (VTTS). VTTS seems to be the next logical step when extending the telephone-centric communication paradigms of yesteryear to the web-centric usage paradigms of tomorrow. In the figure, note that the (audio) TTS system is driving the VTTS adjunct to assure perfect lip synchronization. The synthesized video may be displayed on any device capable of displaying an animated talking head, such as a PC, a PDA (as in Figure 1), or even a display in a car's dashboard.

What are the technological breakthroughs that now make the goal of designing a naturally sounding and naturally looking talking head reachable? One important dimension of the technological advances that may induce paradigm changes like the move towards VTTS and video communications in general is the ever-increasing power of computers. For (audio) TTS, this has a direct effect on the size of the voice inventory we can store and work with. Early concatenative synthesizers (i.e., synthesizers that stitch together snippets of speech to generate an output utterance, e.g., [2, 3, 4]), used very few prototypical units for each class of inventory elements, due to limitations in computational resources. These limitations resulted in what we may consider a "low resolution" representation of the acoustic-phonetic space that a speech synthesizer needs to cover. With a sparse representation of the media space, the problems of distortion and smoothness between

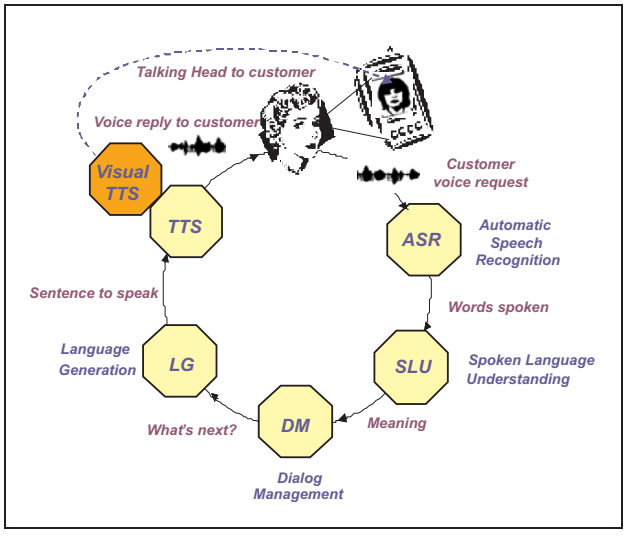

Figure 1. Diagram of any voice-enabled service, extended by Visual Text-to-Speech (VTTS).

concatenated synthesis units become acute. Extensive signal modification (for audio/speech: pitch, amplitude, and duration modifications) are necessary in order to achieve high intelligibility. Unfortunately, such modifications have the potential of destroying naturalness, unless done under very controlled conditions and done very sparingly. More recently, however, the availability of inexpensive but powerful computing resources made it affordable to store many such units, alleviating the need for signal modifications and, therefore, the concatenation and distortion problems [5, 6, 7, 8, 9]. For the synthesis of talking heads (VTTS), the increased storage capacity now available at moderate cost enables storing a larger number of video segments (or extracted face elements such as the moving mouth and the eyes). This is a long shot from the earlier methods of storing just one characteristic frame for a given viseme (the visual equivalent of a phoneme) and using mathematical means to connect these key frames (e.g., [10]). For audio and video TTS alike, the availability of efficient search techniques is also important. These algorithms allow us to search potentially millions of available sound or video units in real time for the optimal sequence that makes up a target utterance. Finally, we now have automatic labelers that can significantly speed up labeling of a voice and face database phonetically and prosodically, making it possible to use voice/face inventories comprising many hours of recorded speech and video. It is important to note that both, the automatic audio (speech) labeler and the optimal search strategy, borrow heavily from the research area of speech recognition (e.g., [11]). This is in addition to the cross-fertilization between TTS and VTTS, where paradigms, algorithms, and tools, are being adapted to address similar problems.

In almost all applications of talking heads as an element of the computer user interface, making a true video recording of a human presenter is a costly proposition. Several applications that require presenting up-to-the-minute information cannot be done with true recordings at all (e.g.,

messaging, database access). In other applications, such as in those running on a customer relations web server, in e-learning, and as a virtual secretary, synthesized lifelike talking heads provide an effective, low-cost solution to the problem of attracting and keeping the attention of viewers/users. In e-learning, for example, research has revealed that talking heads increase the attention span of students [12, 13, 14]. In e-commerce, on the other hand, marketing experts will exert extreme care before letting a synthesized agent represent their product. In addition, and although talking heads can be streamed over the Internet as videos and sound, there is concern about the network bandwidth, latency and server computational load requirements. Here, a talking head that is synthesized on a viewer's terminal equipment (e.g., PC), might be transmitted at a high compression rate, making it viable to stream this kind of audio-visual information even over dial-up modem connections.

Synthesized talking heads may be used to convey non-verbal information in much the same way as a true video recording of a talking person would do [15, 16]. For example, facial expressions are a powerful way to indicate the emotional state of the speaker. And, more generally, virtual agents ("avatars") may be used to direct a user's attention in web navigation [17] and help systems [18, 19], or provide non-verbal information to readers [20].

Avatars have been introduced for some time to represent individual users in Internet chat applications. In general, however, the design process of facial representations in communication or collaboration systems is not straightforward at best, painful at worst [21, 22]. In addition, any user-level tools needed for enhancing avatar-based interactions between human users are still in the early stages of development resulting in less than perfect performance [23].

A recent "hot topic" in user interface design is multimodal interfaces. Such efforts are partially driven by the ubiquitous use of Personal Digital Assistants (PDA's) that largely lack decent keyboards for text input, and partially by the desire for providing an enriched set of modalities for information rendering/output [24]. Researchers found that such enrichment leads to more cooperation from, and to better interaction with, the users.

These results are supported by formal subjective tests aimed at measuring how much a user trusts the computer. It is well known that adding a face to communications increases the trust and cooperation between people as well as between a person and his or her computer. Typically, trust is measured with a classic social dilemma game [25]. In [26], the computer used as part of the interface an animated talking dog, an animated obviously synthetic talking human 3D face model with a texture map, or a life audio and video feed of a real person to interact with the test subject. Using the social dilemma game, the authors found that trust increases from 60% to 78% and 82%, respectively. Obviously, humans adapt their behavior when confronted with a human or human-like representation. Using the same social dilemma game experiments described in [27] that compare the trust for text-only, text with synthetic speech, as well as text with a talking 3D face model with a shaded surface, trust was measured at 52%, 61%, and 67%, respectively. Apparently, human appearance of the animated face is important, too, because the increased trust was not observed when animated faces of dogs represented partners [26]. While humans do not knowingly adapt their behavior, a questionnaire used in conjunction with the experiments in [27] revealed that they expect a higher level of intelligence when the computer is represented with a human-like face in the interface. Given that the trust was highest for the real human in [26], we believe that the use of a face model looking, talking and behaving like a real human will give the highest benefits for the user interface.

Using obviously synthetic talking faces, their importance was also shown in other experiments not specifically targeted at measuring trust. For example, in a typical consumer interview task that traditionally is conducted using paper and pencil questionnaires, the use of a facial display led to fewer mistakes and more time spent on answering questions [28]. In addition, a stern facial expression led to improved responses when compared to a more neutral face. In other research [29], personality traits attributed to the animated face helped users to stay more alert relative to using text-only questionnaires, and led to higher appreciations for personality attributes such as friendliness, cheerfulness, and self-confidence that clearly go beyond simple emotions like joy or sadness [30].

For Visual Text-to-Speech (VTTS), the most aggressive quality goal is to provide computers with synthesized faces that look, talk, and behave like real human faces. Generating lifelike animated faces remains a challenging task despite decades of research in computer animation. To be considered natural, a face has to be not just photo-realistic in appearance, but must also exhibit proper postures of the lips and even the visible portions of the tongue, synchronized perfectly with the speech. Moreover, realistic head movements and emotional expressions must accompany the speech. We are trained since birth to recognize faces, and to scrutinize facial expressions. Consequently, we are highly sensitive to the slightest imperfections in a synthesized facial animation. Only very recently technology has advanced to a point where talking heads can be synthesized with a quality comparable to recorded videos.

A less aggressive (but more practical) goal for VTTS is to match the specific set of features that are required for a given application. In an extreme case, "naturalness" might even be undesirable, for example, when the application designer would like to convey that the user is communicating with a computer, not with a human at the other end of a video connection. Therefore, evaluating the quality of any practical system depends on a careful comparison of desired (planned) and perceived (realized) features [31].

Today, synthesized faces are already an integral part of animated movies and video games, but beyond entertainment, they are slow in finding widespread use. Despite a wealth of data suggesting their potential benefits, only few talking heads appear in commercial applications

such as customer service. Most people associate animated faces with entertainment, and, until recently, the quality of VTTS was not sufficiently high to have synthetic faces act as stand-ins for real humans. This fact had many people question the economic viability of talking heads in 'serious' applications. In the world of down-to-earth business economics, the introduction of new technologies, such as VTTS, into business and consumer services has proven quite difficult. Clearly, more application-oriented testing needs to be done with the focus on what value VTTS adds to applications. Finally, a disclaimer. The authors feel very strongly that the field of quality evaluation of VTTS systems is currently at its infancy, not even close to the still somewhat early stage of evaluating (audio) TTS systems [32]. Therefore, instead of giving the perfect recipe for evaluating VTTS systems, this paper rather aims at highlighting the issues that need to be considered when trying to maximize the quality of a VTTS system. More details of our system can be found in [33].

The remainder of this paper is organized as follows. Section 2 summarizes the methods that we use to create a photorealistic talking head. Section 3 discusses visual prosody that includes facial dynamics that correlate with how the speech is rendered in terms of pitch accents, emphasis, and even emotions. This also includes nods of the head that tend to accompany certain speech events. Section 4 is concerned with quality assessment with a slight bias towards objective means to support algorithmic choices and tuning during development of a VTTS system. Finally, we conclude the paper with section 5.

2. Sample-based Talking-Head Synthesis

Here we summarize the steps we take to synthesize our talking-head animations. As already mentioned, our approach follows the line of the so-called sample-based, image-based, or concatenative synthesis. The basic idea of this approach is to videotape a person uttering a corpus of phonetically balanced sentences. These samples are then prepared and stored into a database so that they can be used with minor processing at synthesis time. The realism of the appearance is due to the fact that the samples are actually recorded units. The challenge of this approach is finding a way to concatenate these samples into a smooth animation while maintaining lip synchronization with the target audio track. In the following, we summarize the steps of sample-based video synthesis: recording, analysis, synthesis and rendering and emphasize the effect on quality of the various parameters used at each steps.

2.1. Corpus Recording

This is arguably the most critical step. Once the technical aspects of synthesizing animations have been sorted out, the quality of the result rests on the initial choice of a suitable talent and the careful recording of the corpus. We direct the talents to speak in a natural way and adjust their speech rate using a teleprompter, thus avoiding over and under-articulation. To improve the quality of the database further, the corpus is typically recorded three times and a manual selection is made to remove unsuitable sequences (speech errors, smiles, smirks, coughs, etc.).

2.2. Image Analysis

Once the selection of usable sequences has been performed, all image frames forming the video sequences are individually analyzed to extract and normalize individual facial features. More details on the techniques used to locate, extract and normalize facial features from images can be found in [34] and [35]. The precision with which these facial parts are extracted is crucial to the quality of the resulting animations. Errors of as little as half a pixel might result in visible artifacts. For example, if, within one particular sequence, for a few frames, the mouth's position is estimated with an offset, these frames, when concatenated with frames from other sequences where the mouth's position was correctly estimated, will produce a sudden, unnatural movement that will be distracting. For this reason, while this step is performed entirely automatically, a final manual inspection of all frames is necessary to ensure a clean database of facial features.

2.3. Unit Selection

As with concatenative speech synthesis [7, 8], we use the Viterbi algorithm to find the best path in a graph representing all possible animations for a given speech target. Note that a unit in the graph is a single video frame. The following costs are assigned to nodes and arcs of the graph: a target cost is assigned to each node and measures the acoustic fitness (phoneme-level) of a particular frame to the given target speech; a concatenation cost is assigned to each arc and measures the visual difference between two consecutive frames, and a skip cost is added to the previous cost that penalizes frame skipping (or, rather, favors keeping frames together as they were recorded). By assigning different weights to these costs, we trade off smoothness against lip-synch. Giving a large weight to the node cost will force a path that closely matches the desired speech target. However, since the node cost does not represent any visual information, the resulting animation might be choppy, being made of many small segments with no visual continuity. Conversely, giving a large weight to the concatenation cost will force a path that is visually continuous, but with less regard to how well the lips are synchronized with the target audio speech. These parameters are typically tuned using the technique described in section 4.1 (a very detailed study of these parameters is presented in [36]).

2.4. Rendering

Audio and face rendering is mandatory for any VTTS system. Both components need to be connected via a synchronization module and a coarticulation engine for creating the correct mouth shapes for the spoken text. Our approach uses the 3D shape and the "background" image sequences of a recorded person's head and shoulder and superimpose

Table I. Some of the ToBI labels marking prosodic events.

Symbol of pitch accent	Movement of the pitch of the fundamental frequency (F0)
H*	High - upper end of the pitch range; typical for accent or stress
H-H%	Pitch high and rising higher towards end; typical for yes-no question
L-H%	Pitch low and rising towards end; typical for comma
L-L%	Pitch low, staying low; typical for end of a statement

on these the optimal sequence of dynamic units selected from a database of normalized and labeled mouth and eye images. Once the script for an animation has been obtained from the unit selection step, the final rendering of the animation is done by overlaying the concatenated animation of mouths onto a 'substrate' face. An alpha channel is used to blend the facial parts progressively into the 'substrate' image of the head to ensure that no hard boundary can be seen between them. The substrate head is typically an entire video sequence of the head, which is selected from the database of recorded video sequences based on matching prosodic characteristics. We describe the process of selecting such 'substrate' video sequences in more detail in section 3.2.

3. Visual Prosody

3.1. Moving the head in synchrony with speech

When we talk, not only the lips and the jaw are moving, but usually the whole head moves as well, and often gestures and movements of the whole body accompany the speech. Moreover, the face may exhibit expressions in synchrony with the spoken text. Some of these movements are intentional and related to the meaning, while others do not have an obvious connection with the content of the speech [37]. Facial expressions related to speech have been studied extensively in the psychology literature, but there exist, to our knowledge, no studies describing amplitude and duration of these events in a quantitative way that could be used for driving animations. For naturally appearing animations, head and facial movements are critical and a lack thereof gives the talking head a synthetic look [38]. Important for our purposes is that head movements can be inserted based on prosodic information only, and do not need a semantic interpretation of the spoken text. We know from audio-only TTS that it is relatively easy to determine the prosody from a given text automatically, while extracting its meaning is unreliable. Prosody prediction is a major task for Text-to-Speech synthesizers and well-developed tools exist to execute this reliably.

In order to get quantitative descriptions of head movements we analyzed several hours of recorded video, extracting prosodic phrase boundaries and pitch accents, as well as the precise movements of the head, using the image analysis tools mentioned above. Details can be found in [39].

Prosodic events are commonly labeled according to the ToBI (Tones and Break Indices) prosody classification scheme [40]. ToBI labels do not only mark accents and boundaries, but also associate them with a symbolic description of the pitch movement in their vicinity. Table I shows ToBI symbols indicating the movement of the fundamental frequency (F0). The two-tone levels, high (H) and low (L), denote the pitch relative to the local pitch range and baseline.

3.2. Prosodic head movements

For identifying prosodic head movements, the three angles of rotation, together with the three translations are measured. For this analysis, each of the six signals representing rotations and translations of the head is high-pass filtered, eliminating components below 2 Hz. Movements in the low frequency range extend over several syllables and often over multiple words. Such movements tend to be caused by a change of posture of the speaker, rather than being related to prosodic events in the speech. The faster movements, on the other hand, tend to be closely correlated with prosodic events. Accents are often underlined with nods, extending typically over two to four phones, as can be seen in Figure 2.

In order to obtain a compact representation of these movements we classify them into a small number of characteristic motion patterns:

- Nod, i.e. an abrupt swing of the head with a similarly abrupt motion back.
- Nod with an overshoot at the return, i.e. the pattern looks like an 'S' lying on its side.
- Abrupt swing of the head without the back motion. Sometimes the rotation moves slowly, barely visible, back to the original pose, sometimes it is followed by an abrupt motion back after some delay.

This basis set of motion primitives provides a simple framework for categorizing and describing head movements quantitatively with amplitudes and durations. Figure 2 shows how the measured movements are binarized, based on their derivatives with respect to time. Such curves are then used to classify motion patterns.

Head and facial movements during speech exhibit a wide variety of patterns that depend on personality, mood, content of the text, and other factors. Yet, while angles and amplitudes of the head movements exhibit large variations, their rhythm tends to be strongly correlated with the prosodic events of the text. The same is true for rises of eyebrows that are often placed at prosodic events, sometimes in combination with head nods, at other times without. Visual prosody is not as predictable as acoustic prosody, but is clearly identifiable in the speech of most

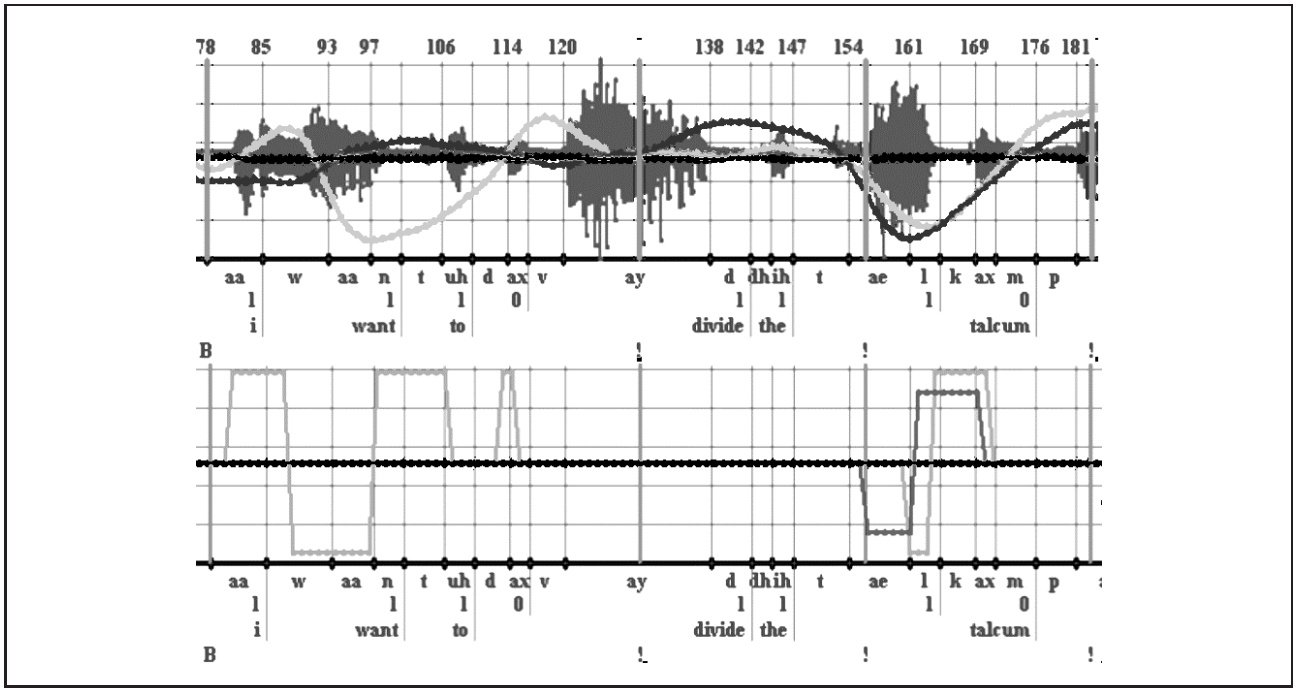

Figure 2. Head movements; rotations around x (light gray), y (medium gray), and z (black) axis as a function of time. The thicker vertical lines mark the start of the sentence and the stresses. Also shown is the acoustic signal of the sentence "I want to divide the talcum ..." At the top of the charts are the frame numbers; the phonemes and word boundaries are indicated below. The top image shows the measured values. In the bottom figure, the derivatives with respect to time have been binarized (clipped) in order to identify basic motion patterns. These patterns are used to measure similarities of visual prosody.

people. Table II shows the number of prosodic events and of accompanying head movements in a corpus of 20 minutes of spoken text.

A basis set of motion primitives plus the statistics of frequency and amplitude of these movements provides a framework for driving a talking head. Different personalities can be emulated by using the statistics of a particular person. For our sample-based approach, we use recorded segments of prototypical head movements that are concatenated. This can be done with a relatively small number of sample movements, since the timing of these movements does not have to be very precise. In order to find matching sequences for a new sentence, the prosodic events produced by the TTS are compared to those of recorded sentences in the database. We assume that if the acoustic prosody is similar, the visual prosody from the recorded sentence is going to be a good match for the new sentence. Hence, we concatenate recorded sequences of head movements that match the new sentence in prosody and then overlay the mouth shapes onto these background heads. This results in animations that look very natural and where it is often difficult to decide whether they were synthesized or recorded.

3.3. Evaluating the effect of visual prosody

We conducted extensive experiments with different types of head movements in order to determine what is perceived as natural. The main result is that the movements of the head have to be synchronized with the speech in order to look convincing. We generated animations with and without head movements and also compared animations where the head moved randomly with animations where head movements were placed at prosodic events. Invariably the animations with well-synchronized movements were preferred. Many viewers made the comment that with these movements the head appears to understand what it says, while without them the head appears disengaged and robot-like. Table III summarizes the results of a test conducted with 22 viewers, each of them judging 5 sentences that were synthesized with random head movements, prosodic movements, and prosodic movements with larger amplitudes. The large increase of a full MOS point (Mean Opinion Score) score from random to prosodic movements underlines the significance of well synchronized head movements. The movements synthesized for the third part of test, marked 'stronger prosodic movements' were taken from a database of recordings where the speaker was instructed to act 'happy'. In this database, the head

Table II. Number of prosodic head movements in 20 minutes of spoken text (300 sentences), recorded with one person. In this case roughly one third of all prosodic events are accompanied by a characteristic head motion, such as a nod. PTE: Predicted Textual Events; MVE: Matching Visual Events.

	PTE	MVE
Beginning of Sentence	299	97
Accent	1158	423
End of Sentence	296	62

Table III. In these test results: 22 viewers were judging 5 sentences synthesized with 3 different types of head movements. They were asked, "How natural does the head look" and gave their opinion on a 0 to 5 scale; the average of these numbers is the 'Mean Opinion Score' (MOS) shown here. (see text for details about the head movements). h.m.: head movements, p.m.: prosodic movements.

Random h.m.	Synchronized p.m.	Stronger p.m.
MOS: 2.8	MOS: 3.8	MOS: 3.5

movements were then more pronounced, roughly +-6 degrees for the nods, versus about +-3 degrees when the speaker was instructed to act 'neutral'. Viewers judged these stronger head movements as less natural than the more subtle ones. When asked about this difference in perception, some viewers remarked that the stronger movements looked like 'overacting' by the speaker. This indicates that the strength of the movements should be adjusted to the situation. In order to convey happiness or liveliness, stronger head movements may be appropriate, while for neutral sequences more subtle movements are in order.

4. Quality Assessment

Determining the quality of an animation in an objective way is quite difficult, because no universally accepted criteria exist how a talking head is supposed to look. Evaluations of intelligibility typically assess lip-reading capabilities [41, 42, 43] in a subjective test. The standard approach to assessing naturalness for VTTS is also to conduct subjective tests where human observers provide feedback on a scale from 0 to 5, resulting in Mean Opinion Scores (see also Table III) [44]. However, this is time consuming and expensive, since we need a large number of observers, preferably from different demographic groups, to look at a large number of samples. Therefore, we established a few measurable features whose values strongly correlate with the quality assessments given by human observers. This technique greatly accelerates development because it provides instantaneous feedback whether changes in the synthesis algorithms result in better quality. Such an automatic quality assessment will not replace subjective tests completely, but can greatly reduce the need for them.

In order to establish a reference against which the automatic quality assessments can be calibrated, a set of 50 sentences was synthesized and viewers provided subjective scores for each of them. Thirty viewers expressed their opinion about 'synchronicity', 'smoothness', and 'precision' of the animations. All sentences had also been recorded, so that we could compare directly synthetic animations and recorded videos. For these tests all synthetic animations used the recorded sound track. In this way, the viewers were not distracted by imperfections of the sound and could focus their attention exclusively on the visual effects. Each viewer judged around 20 sentences and provided results on the 0 to 5 MOS scale. The viewers were not educated about the technical aspects used to create the animations, and many saw such animations for the first time. Test sequences were presented in a laboratory environment where viewers were invited for sessions of about 15 minutes in length. Alternatively, the tests were shown on laptops placed in office environments with the sound coming from the built-in speakers. On the screen the talking heads were the only graphic elements present - two or three heads side-by-side that the viewers were asked to compare.

The absolute values provided by viewers are typically not meaningful, and therefore tests have to be designed for direct comparisons between different samples. Viewers are instructed that on the scale of 0 to 5, 0 means 'very bad' or 'poor' and 5 is 'very good' or 'excellent'. Inevitably, different people have different opinions about what is good or bad, and consequently, when asked if a sample is 'natural', for example, the scores may vary widely. Relative scores, however, judging whether one sample is better than the other, are much more consistent, at least in their sign, but not necessarily in their absolute values. Hence, all these tests were side-by-side tests, primarily used to answer questions of the type: "If we change this parameter in the synthesis, is the perceived quality (synchronicity, smoothness, etc.) better or worse". Results were evaluated by calculating the mean of all answers, as well as the statistically more robust median, and discrepancies between the two helped identify outliers. A face triggers immediately a variety of reactions in a viewer. For example, the viewer may not like the hairstyle or attribute such traits as 'friendly' or 'arrogant' to the face. Some people focus more on the eyes than on the mouth and, when asked about the reason for their (low) scores, may mention 'staring eyes' first, rather than a quality related to the mouth. Hence, we chose three quality criteria, 'smoothness', 'synchronicity' and 'precision' (see below), that are clearly related to the articulation, and force the viewer to focus on the mouth rather than judging the whole face. People need some instructions in order to be able to judge these criteria, but the resulting scores are easier to interpret and more informative for our purposes than scores produced for such questions as 'how natural does the face look?'.

4.1. Automatic quality evaluation: Comparing recorded and synthetic sequences

The easiest and most reliable way of measuring the quality of an animation is to synthesize a sentence that has also been recorded and compare the two frame by frame. An example of such a comparison is shown in Figure 3, where the mouth height is depicted for a recorded and a synthesized utterance. For the animation, the recorded sound was used in order to guarantee that the timing of the phones is the same in both cases.

By comparing the quality scores given by viewers with measurable parameters, several features were identified that have a large influence on the perceived quality of the articulation:

- Presence of lip closures, openings and protrusions,

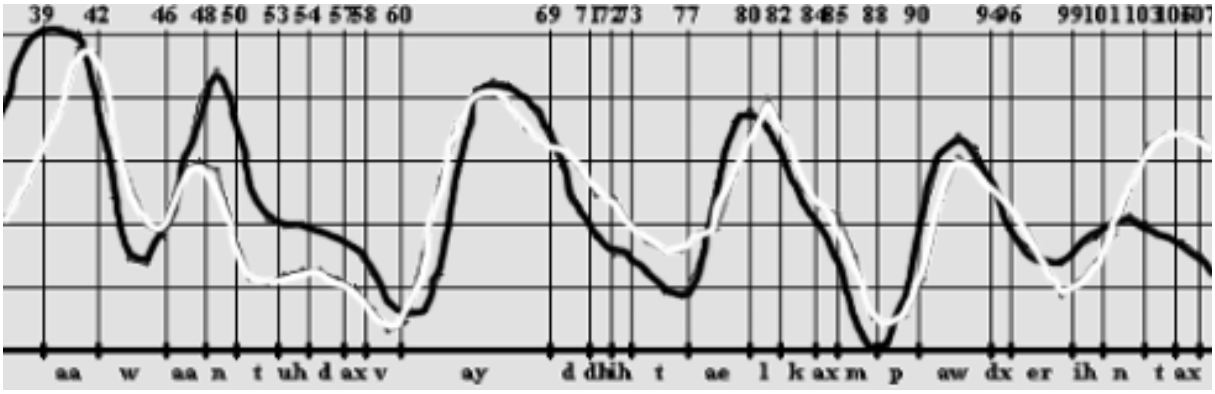

Figure 3. Mouth height as a function of time for the sentence: "I want to divide the talcum powder into . . . ". The black line shows the recorded values, the white one the values of the synthesized articulation. For a quality score of 'precision', the differences in mouth height between recorded and synthetic samples are summed and normalized by the number of frames.

- Timing precision of closures, openings and protrusions,
- Timing precision of turning points (points where lip direction changes from opening to closing and vice versa),
- Smoothness.

Figure 4 shows a short example of the mouth height versus time, where the turning points are marked. Moreover, also shown are the thresholds defining whether the mouth is closed or open. Here, 'open' really means wide open, as for example for the vowel /ah/ in 'but'. What is considered 'open' varies considerably from one person to another. Many people articulate rather discretely and their 'wide open' mouth height may be much less than what we see for people who articulate strongly. Nevertheless, even for those people who do not open widely, it is important to show clear mouth openings and place them precisely. Otherwise, the animation gives the impression of slurred speech. The parameters used for these measurements, such as 'wide open' or 'closed' are determined individually for each recorded person. For example, the thresholds for 'wide open' are determined by measuring the maximum mouth height for /ah/ in 100 sentences and then taking 75% of the mean value as threshold. Similarly, the presence of a protrusion is detected by setting a threshold, based on the minimum mouth widths measured during the articulation of /uh/.

Closures have an important effect on the perceived quality. These events are clearly visible in every person's articulation, since for several consonants, such as plosives and bilabials, the lips have to be closed. A lack of closures gives the impression that the lips and the sound are 'disconnected', i.e. the lips are not really articulating the text that is spoken. Protrusions of the lips are also clearly noticeable events that have to be placed precisely for good quality. However, their effect on quality is not quite as pronounced as that of the closures. In fact, there are speakers who show minimal lip protrusions in their articulation. Nevertheless, the presence of clear protrusions placed precisely enhances the perceived quality of the animations considerably. Missing a closure every now and then does not seem to subtract much from the subjective quality score, but once more than 20% of the closures are missed a clear degradation of the quality score is observed. For protrusions, their presence in at least 50% of the cases seems sufficient. It has to be understood that the objective measures depend on hard thresholds and even if the automatic scorer is not recording a protrusion, there may still be a subtle narrowing of the lips present that is interpreted as slight protrusion by the viewer.

'Smoothness' is also quite critical for achieving good quality. The lack thereof is perceived as jerkiness in the articulation and is often irritating to a viewer. Viewers typically have difficulties with coming up with explanations why they perceive articulation as jerky, but it is one of these characteristics that "I know it when I see it". For the objective measure, smoothness is determined by measuring differences in mouth height in neighboring frames. However, where in the text this difference is measured is also important. For example, a transition from a plosive to a wide open vowel, such as /p-ah/ results in a very fast opening of the lips and does not appear as jerky. Similar opening speeds in other parts of an utterance may be perceived as unnatural.

Based on these observations, we developed a quality score for 'synchronicity' that takes into account presence and placement precision of closures, protrusions, openings, and turning points. The distances of these events in the animated articulation from the ones in the recorded video are summed and then normalized to provide the 'synchronicity' score (compare Figure 4). For 'precision', the differences in lip width and height of the synthesized and recorded articulation are summed (compare Figure 3). For the 'smoothness' score, the differences in mouth heights across segment boundaries are added. Table IV shows a comparison of objective, automatic quality scores and subjective scores, obtained from eight human observers. In order to judge the precision, the viewers were shown recorded and synthesized versions of the same sentences and asked to judge how precisely they match. They

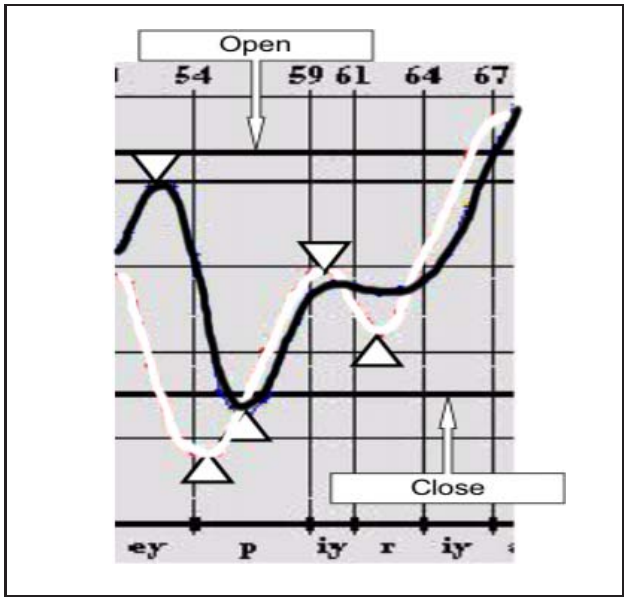

Figure 4. Mouth height as a function of time for a recorded sequence (white) and a synthesized one (black). Marked are the turning points, where the motion changes from opening to closing and vice versa. The visual impression of well-synchronized articulation depends on the locations of these turning points. For the objective measure of 'synchronicity' the average time differences between corresponding extrema in the recorded and synthetic sequences are determined.

Table IV. Comparison of automatic and subjective score for a test set of 20 sentences; 8 people were evaluating the sentences for the subjective test. The quality is judged on an MOS scale from 0 (poor) to 5 (excellent). Subjective scores were measured in recorded-synthetic side-by-side tests and scores for the recorded samples are normalized to 5. The automatic scores were calibrated by setting automatic scores to subjective ones for 10 sentences (training set) that were not part of the 20 sentences scored for the test shown here (test set). Note that perceived smoothness is content-dependent (large sudden openings/closings are tolerated in some circumstances) and hence does not always match the calculated one.

	Automatic Score	Subjective Score
Precision	4.61	4.62
Synchronicity	4.39	4.78
Smoothness	2.48	4.29

were also asked how well the sound is synchronized with the lip articulation (synchronicity) and how smooth the articulation is (smoothness). The observers judged these criteria on the MOS scale of 0 (bad) to 5 (excellent). For all the subjective scores the viewers were shown recorded and synthetic sequences side-by-side. In order to enable comparison with objective scores, subjective scores for the recorded samples were "shifted" to (anchored at) a value of 5, and the same "shift" was applied also to the scores for the synthetic samples (i.e., maintaining only the differences in scores). The proper normalization for the objective metrics was determined by comparing subjective and objective quality scores of 10 sentences (that were not part of the 20 test sentences). As can be seen in Table IV, precision and synchronicity tend to agree well with those of human observers. Smoothness is not captured that well with our criteria. This shortcoming is due to the fact that, when measuring differences in mouth height, the position within the text is not taken into account. This relates back to the speed and position issues mentioned in the previous paragraph. Clearly, more work needs to be done to find a better objective measure related to perceived smoothness. Note, however, that perceived smoothness is content-dependent (large sudden openings/closings are tolerated in some circumstances) and hence does not always match the calculated measure.

4.2. Online Quality Evaluation

When quality has to be assessed for new sentences that have never been recorded, we cannot rely on comparisons between synthetic and recorded articulation. We developed several quality criteria that can be used to judge synthesized articulations alone. One quality parameter that can be measured easily is the speed of mouth opening or closing. As mentioned in section 4.1, it is important to take the type of phoneme into account when judging the speed of mouth opening and closings. From the recorded database, we establish maximal values for the change in mouth height between neighboring frames. Some diphone transitions, such as /p-ah/, show a very rapid mouth opening, while for others, such as /n-ah/, the lip movements are typically much less pronounced. The recorded database was scanned for transition speeds and maximal values were established for several groups of diphones. These values were then used to determine whether in a synthesized sequence the transitions are within the tolerances. The number of frames where these boundary values are exceeded gives a good criterion for judging the 'choppiness' of an animation.

Another quality parameter is the number of missed closures that is measured by comparing the mouth height to a threshold value (compare Figure 5). Moreover, the precision of the closures is determined by measuring the distance (in frames) of the minimal mouth height from the predicted closure position based on phoneme positions and durations. Similarly, the presence of a protrusion at the appropriate places is determined. In addition to the parameters that relate directly to visual effects, we also take the average length of contiguous segments from the database as a quality criterion. Experience has shown that if the average length of recorded segments concatenated for the articulation is five frames or more, the articulation is likely to appear smooth. Typically the quality score for 'smoothness' is low if the average recorded segment length is less than 5 frames and improves for average segment lengths up to 10. At that length the score for 'smoothness' is approaching the score of recorded sequences.

4.3. Passing the Turing Test

Our stretch goal is to produce animations that pass the Turing test, namely that a viewer cannot distinguish between

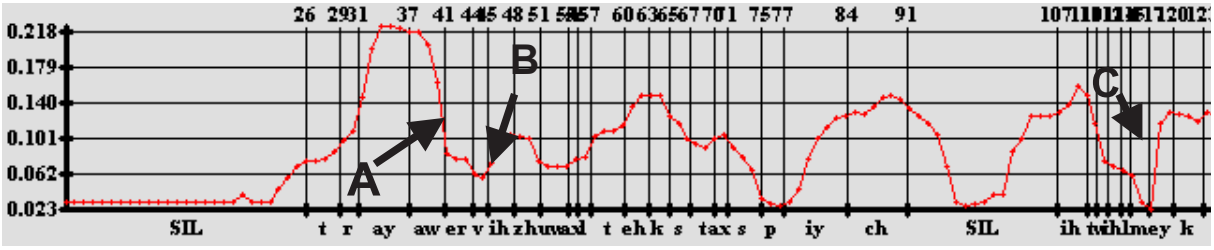

Figure 5. Mouth height as a function of time for the sentence: "Try our Visual Text-to-Speech, it will make . . . "; The arrows show where the automatic quality evaluation flagged errors: A: closure too fast; B: missed closure; C: opening too fast. The average contiguous segment length is 6.05 frames with a variance of 2.4 frames, a value judged as good (main influence on 'smoothness').

animation and recording. For several sequences that were synthesized, this has actually been achieved. For this test 25 viewers were asked to judge two sentences side-by-side, one was the recorded sentence and the other one a synthetic sentence. In either case, the recorded audio was used, otherwise the TTS would give it away immediately which sentence is synthetic. Out of a test set of 10 synthesized sentences, 3 were marked 'recorded' with the same probability as the truly recorded ones. Five were judged synthetic by about two thirds of the viewers while one third thought they were recorded and two had an obvious artifact that made it clear that they were synthesized. The sentences presented for this test were taken from the database recorded to cover most of the English diphones. They were statements, 10 to 20 words long, such as, for example, "I want to divide the talcum powder into two piles".

A Turing test is much more stringent than what is required for presenting convincing animations for such applications as customer service, since the direct comparison of synthetic and recorded sequences lets the viewer notice even slight differences that otherwise would be missed. When viewers were asked to judge if a sequence is 'natural' in a random selection of synthetic and recorded sequences, 80% of the synthesized ones had a score similar to the recorded ones. However, it has to be emphasized that such close similarities with recorded sequences is achieved only for relatively short sentences. When long sentences and whole paragraphs are synthesized, the behavioral patterns of the head make it obvious whether it is synthetic or recorded. Hence when judging the quality of animations, several levels of difficulty have to be distinguished:

1. Short speech articulation, no other movements,
2. Short articulations with prosodic movements,
3. Longer articulation with prosodic movements and emotions,
4. Articulation with prosodic movements, emotions and behavioral patterns.

This approach is very much in line with what has been suggested in audio-only TTS, where we are currently able to pass the Turing test for short utterances and further research is aiming at passing it for more complex utterances [45].

For levels 1 and 2 the present technology can generate animations that are of a quality comparable to recorded videos. A formal test is reported in [46] where morphed articulations were compared to recorded videos. In one test, the viewers were shown an animation and asked to judge whether it is real. In a second test animation and recorded video were shown side-by-side. In both cases, the animations were indistinguishable from recordings. Such video-realistic animations have been achieved only with sample-based techniques. To our knowledge, so far, no model-based 3D heads have shown articulations with a quality comparable to recordings.

Levels 3 and 4 require a semantic interpretation of the text in order to introduce emotional expressions and behavioral patterns that appear meaningful. This is beyond present day natural language understanding and, for the time being, this requires interpretation by humans who then annotate the text with tags. Natural appearance can, in principle, be achieved with sample-based techniques. In practice, however, this may require such large databases of recorded samples that it may not be feasible beyond emulating some basic behaviors.

5. Conclusions

Visual Text-to-Speech (VTTS) synthesis has come a long way towards producing high quality synthetic output. Following the earlier lead towards higher quality in audio-only TTS, VTTS is now opening the possibility of generating synthetic "talking heads" of such a quality that they may be mistaken for recordings of real humans.

Standard components of any VTTS system are an audio and face renderer connected via a synchronization module and a coarticulation engine for creating the correct mouth shapes for the spoken text. Our approach to VTTS is to use the 3D shape and the "background" images sequences of a recorded person's head and shoulder and superimpose on these the optimal sequence of dynamic units selected from a database of normalized and labeled mouth and eye images. We found that for animations of high quality, the database must contain tens of thousands of sample textures resulting in a memory footprint of the VTTS system that easily exceeds 100 Mbytes, using a talking head at 256 by 256 pixel resolution. However, this size can be reduced to less than 2 Mbytes by using more aggressive compression and a reduced video size of 100 by 100 pixels. Following

earlier findings in work on audio-only TTS, where extensive "warping" of units by signal processing led to lower naturalness, we currently try to cover all appearances with recorded samples, thus incurring minimal or no deformation of the textures.

We found that using head and eye movements that correlate with events in the corresponding speech (e.g., stress pattern of a sentence) contribute significantly to perceived naturalness. Again, this finding is consistent with results in audio-only TTS, where prosody, the way how a sentence is spoken, is an important determinant for perceived naturalness. Consequently, we also included in our VTTS system a prosody analyzer, and a visual prosody generator. Both components are responsible for creating the appropriate eye and head motion, based on the prosody of the spoken text. However, much more work remains to be done to analyze and synthesize appropriate prosodic movements of humans in various situations. For example, head movements have been analyzed previously either qualitatively or with motion capture equipment. However, little has been done so far to categorize prosodic movements, and appropriate models do not exist at present. Hence, it still remains to be seen where the optimal trade-offs are between the use of models and the use of direct data such as recorded sequences of background face images.

As quality improves, evaluation paradigms that help drive this process are critical. Clearly, subjective tests are essential for judging visual prosody and the overall effectiveness of using VTTS in human-computer interactions. However, we have also developed objective metrics for determining the quality of a talking mouth based on criteria like mouth closures, mouth protrusions, turning points in lip motion direction and motion smoothness. These metrics helped us tremendously with making the right trade-offs between quality and database size, for example, and with algorithmic choices in unit selection cost measures.

For an interactive demonstration of our sample-based talking-head synthesizer, we encourage the reader to visit our web site at `http://vir2elle.com`.

References

[1] R. V. Cox, C. A. Kamm, L. R. Rabiner, J. Schroeter, J. Wilpon: Speech and language processing for next-millennium communication services. Proc. IEEE, 88, August 2000, 1314–1337.

[2] D. O'Shaughnessy, L. Barbeau, D. Bernardi, D. Archambault: Diphone speech synthesis. Speech Communication **7** (1988) 55–65.

[3] R. Sproat, J. Olive: Text to speech synthesis. AT&T Technical Journal **74** (1995) 35–44.

[4] R. Sproat, J. Olive: Text-to-speech synthesis. – In: The Digital Signal Processing Handbook. V. K. Madisetti, D. B. Williams (eds.). CRC Press, IEEE Press, 1998, Ch. 46.

[5] Y. Sagisaka, N. Kaiki, N. Iwahashi, K. Mimura: ATR–ν-TALK speech synthesis system. Proc. Int. Conf. on Speech and Language Processing 92, Banff, Canada, 1992, vol. 1, 483–486.

[6] A. W. Black, P. A. Taylor: CHATR: A generic speech synthesis system. COLING '94, 1994, 983–986.

[7] A. Hunt, A. Black: Unit selection in a concatenative speech synthesis system using a large speech database. Proc. ICASSP **1** (1996) 373–376.

[8] A. D. Conkie: Robust unit selection system for speech synthesis. Joint Meeting of ASA, EAA and DAGA, Berlin, Germany, 15-19 Mar, 1999, paper 1PSCB_10.

[9] S. Deligne, F. Yvon, F. Bimbot: Selection of multiphone synthesis units and grapheme-to-phoneme transcription using variable-length modeling of strings. – In: Data-Driven Techniques in Speech Synthesis. R. I. Damper (ed.). Kluwer Academic Publishers, 2001, Chapter 6.

[10] T. Ezzat, T. Poggio: MikeTalk: A talking facial display based on morphing visemes. Proc. IEEE Computer Animation, 96-102, 1998.

[11] M. Ostendorf, I. Bulyko: The impact of speech recognition on speech synthesis. Keynote paper. Proceedings IEEE 2002 Workshop on Speech Synthesis, Santa Monica, Sept. 11-13, 2002.

[12] W. L. Johnson, J. W. Rickel, J. C. Lester: Animated pedagogical agents: Face-to-face interaction in interactive learning environments. International Journal of Artificial Intelligence in Education **11** (2000) 47–78.

[13] C. Elliott, J. Rickel, J. Lester: Lifelike pedagogical agents and affective computing: An exploratory synthesis. – In: Artificial Intelligence Today, Lecture Notes in Computer Science 1600. M. Wooldridge, M. Veloso (eds.). Springer-Verlag, 1999, 195–212.

[14] J. Lester, S. G. Towns, C. B. Callaway, J. L. Voerman, J. FitzGerald: Deictic and emotive communication in animated pedagogical agents. – In: Embodied Conversational Agents. J. Cassell, J. Sullivan, S. Prevost, E. Churchill (eds.). MIT Press, Boston, 2000, 123–154.

[15] E. Andre, T. Rist, J. Muller: Guiding the user through dynamically generated hypermedia presentations with a lifelike character. Proc. Intelligent User Interfaces, 1998, 21–28.

[16] T. Rist, E. Andre, J. Muller: Adding animated presentation agents to the interface. Proc. Intelligent User Interfaces, 1997, 79–86.

[17] A. E. Milewski, G. E. Blonder: System and method for providing structured tours of hypertext files. US Patent # 5760771, June 2, 1998.

[18] A. Don, T. Oren, B. Laurel: Guides 3.0. Video Proc. ACM CHI, 1993, 447–448.

[19] S. Gibbs, C. Breiteneder: Video widgets and video actors. Proc. UIST, 1993, 179–185.

[20] T. Bickmore, L. Cook, E. Churchill, J. W. Sullivan: Animated autonomous personal representatives. Proc. Intl. Conf. on Autonomous Agents, 1998, 8–15.

[21] J. R. Suler: From ASCII to holodecks: Psychology of an online multimedia community. Presented at the Convention of the American Psychological Association, Chicago, 1997.

[22] I. S. Pandzic, T. K. Capin, E. Lee, N. Magnenat-Thalmann, D. Thalmann: A flexible architecture for virtual humans in networked collaborative virtual environments. Proc. Eurographics **16** (1997) 177–188.

[23] B. Damer, C. Kekenes, T. Hoffman: Inhabited digital spaces. Proc. ACM CHI, 1996, 9–10.

[24] M. Johnston, S. Bangalore, G. Vasireddy, A. Stent, P. Ehlen, M. Walker, S. Whittaker, P. Maloor: MATCH: An architecture for multimodal dialogue systems. Proceedings of the 40th Annual Meeting of the Association for Computational

Linguistics, 2002. Available at: http://www.research.att.com/∼johnston/matchacl02.pdf.
[25] R. Axelrod: The evolution of cooperation. Basic Books, New York, 1984.
[26] S. Parise, S. Kiesler, L. Sproull, K. Waters: My partner is a real dog: Cooperation with social agents. Proc. CSCW, 1996, 399–408.
[27] J. Ostermann, D. Millen: Talking heads and synthetic speech: An architecture for supporting electronic commerce. Proc. ICME, 2000, MA2.3.
[28] J. H. Walker, L. Sproull, R. Subramani: Using a human face in an interface. Proc. ACM CHI, 1994, 85–91.
[29] L. Sproull, M. Subramani, S. Kiesler, J. H. Walker, K. Waters: When the interface is a face. Proc. Human–Computer Interaction **11** (1996) 97–124.
[30] J. Ostermann: E-COGENT: An electronic convincing agent. – In: MPEG-4 Facial Animation: The Standard, Implementation and Applications. I. S. Pandzic, R. Forchheimer (eds.). Wiley, Chichester, England, 2002, 253–264.
[31] U. Jekosch: Sprache, Hören und Beurteilen: Ein Ansatz zur Grundlegung der Sprachverständlichkeitsbeurteilung. Habilitation thesis (unpubished, in German), University of Essen, Germany, 2000.
[32] J. P. H. van Santen, L. C. W. Pols, M. Abe, D. Kahn, E. Keller, J. Vonwiller: Report on the 3rd ESCA TTS workshop evaluation procedure. Third ESCA Workshop on Speech Synthesis, Jenolan Caves, Australia, 1998.
[33] E. Cosatto, J. Ostermann, H. P. Graf, J. Schroeter: Lifelike talking faces for interactive services. Invited Paper, Proc. of the IEEE, Special Issue on Human-Computer Multimodal Interface **91** (2003) 1406–1429.
[34] H. P. Graf, E. Cosatto, T. Ezzat: Face analysis for the synthesis of photo-realistic talking heads. Proc. Fourth IEEE Int. Conf. Automatic Face and Gesture Recognition, Grenoble, France, IEEE Computer Society, Los Alamos, 2000, 189–194.
[35] H. P. Graf, E. Cosatto, G. Potamianos: Robust recognition of faces and facial features with a multi-modal system. Proc. IEEE Systems, Man and Cybernetics, 1997, 2034–2039.
[36] E. Cosatto. These No 2675, Ecole Polytechnique Federale Lausanne, Lausanne, 2002.
[37] J. B. Bavelas, N. Chovil: Faces in dialogue. – In: The Psychology of Facial Expression. J. A. Russell, J. M. Fernandez-Dos (eds.). Cambridge U. Press, Cambridge, 1997, 334–346.
[38] J. Beskow, B. Granström, D. House: A multi-modal speech synthesis tool applied to audio-visual prosody. – In: Improvements in Speech Synthesis - COST 258: The Naturalness of Synthetic Speech. E. Keller, G. Bailly, A. Monaghan, J. Terken, M. Huckvale (eds.). John Wiley and Sons, 2002, Ch. 38, 22–38.
[39] H. P. Graf, E. Cosatto, V. Strom, F. J. Huang: Visual prosody: Facial movements accompanying speech. Proc. Int. Conf. on Automatic Face and Gesture Recognition, 2002, 396–401.
[40] K. Silverman, M. Beckman, J. Pitrelli, M. Ostendorf, C. Wightman, P. Price, J. Pierrehumbert, J. Hirschberg: TOBI: A standard for labeling English prosody. Proc. ICSLP, Banff, 1992, Vol. 2: 981–984.
[41] D. W. Massaro: Perceiving talking faces: From speech perception to a behavioral principle. MIT Press, Cambridge, Massachusetts, 1998.
[42] I. S. Pandzic, J. Ostermann, D. Millen: User evaluation: synthetic talking faces for interactive services. The Visual Computer Journal, Springer Verlag **15** (330-40) 1999.
[43] J. J. Williams, A. K. Katsaggelos: An HMM-based speech-to-video synthesizer. IEEE Trans. on Neural Networks, Special Issue on Intelligent Multimedia **13** (July 2002).
[44] ITU-R BT.500-10 1: Methodology for the subjective assessment of the quality of television pictures. 2000.
[45] J. Schroeter, A. Conkie, A. Syrdal, M. Beutnagel, M. Jilka, V. Strom, Y. J. Kim, H. G. Kang, D. Kapilow: A perspective on the next challenges for TTS research. IEEE Signal Processing Workshop on Speech Synthesis, Santa Barbara, CA, September 2002.
[46] T. Ezzat, G. Geiger, T. Poggio: Trainable videorealistic speech animation. Proc. ACM SIGGRAPH, 2002, 388–397.