Robots with Bad Accents
Living with Synthetic Speech

Submitted December 2006, Revised March2007

Vol. 41, No. 3, Leonardo, MIT Press, Cambridge, Massachusetts

Marc Böhlen,
Artist, Engineer
MediaRobotics Lab, Department of Media Study, University at Buffalo
AILAB, University of Zurich
marcbohlen@acm.org

ABSTRACT

Synthetic speech technologies have a profound impact on how we think about and interact with computers. This text discusses parts I and II of the 'Make Language Project', a trilogy on the cultural fallout of machine generated speech as a conduit for reconsidering prejudices in synthetic speech production and humanoid robot design.

Normalized Android Culture

Born from the industrial revolution's promise for a life of plenty and leisure, robotics is firmly committed to the positive, utopian interpretation of technology as first formulated by early thinkers such as Moore, Spencer, Saint-Simon, and reinterpreted in terms of computer technology in the 20th century by the cybernetics' community [1]. Not everyone shared this view. The sociologist Sorokin [2], for example, imagined human advancement through technology would end in disaster. The odd contradictory mix of awe, angst and admiration with which high-end robots are perceived today is proof of the continued vigor of the polarized view points; the intellectual landscape seems firmly settled with engineers and scientists on the positivist side, and humanities scholars and artists mostly on the pessimistic side, with some interesting scholars suggesting a compromise, as it were, by claiming the future of technology to end in utter uselessness [3].

From Turing to Kurzweil and beyond into popular culture [4], the capacity to recall more and calculate faster has been directly associated with super-human intelligence. Because the illusive goal of superior intelligence is not practically achievable, research agendas have concentrated on matching human intelligence and behavior in select domains. Not surprisingly, even this less lofty goal is far from trivial. Computational vision, for example, is still struggling to achieve synthetic visual perception and processing on par with that of humans. Likewise, the field of humanoid robotics does not currently attempt to make machines superior to humans; rather it has moved its focus to devices and processes that mimic humans. Interestingly, this notion of similarity or equality is defined in very specific ways and along strong disciplinary assumptions and rhetorical goals. For example, as Nourbakhsh and others have observed, most robots are designed as pets or servants [5], and they are all benevolent and polite [6]. Furthermore, humanoid and android robot designers tend to recreate physical perfection in their products. Ishiguro [7], for example, used an attractive young television moderator as a model for his most advanced and uncomfortably realistic android.
Despite the immediate appeal, beauty, benevolence and politeness are problematic machine design

guidelines. They normalize android culture and create a sympathetic base for robots that the machines do not necessarily deserve. By normalizing android culture, one looses opportunities for interaction forms that are uncomfortable and problematic but, potentially, rich and complex. Normalized android culture leaves us the promise of a friendly utopia that might well remain unfulfilled; it promises a future that is only superficially friendly, and leavs us unprepared to deal with conflicts that will likely arise with sentient machines in the future.

Confrontational Interaction

Normalizing machines to behave as humans do in select social contexts limits the scope of research in robot design. It also creates a fragile and shallow basis for any kind of deep exchanges between robots and people that the social robotics agenda claims to address. But if deep and longterm exchanges between synthetic systems and real people are to be achieved, a wider basis for possible forms of exchange and ways of sharing between machines and people is of essence. Synthetic systems, complex, confused and contradictory as we humans are will make, over time, for better partners than polite drones. Ultimately, the goal is diversity in robot design, a diversity defined not only in technical terms, but also by varied ideas about what machines could be and what we can share with them. In this context I offer some preliminary observations from the 'Make Language' trilogy [8], parts I and II, in which synthetic accented speech and machinic foul language infringe on comfort zones of interaction with computational devices.

Text To Speech Synthesis

Humans are uniquely specialized in the production of speech, and only homo sapiens can use tongue, cheeks, lips and teeth to produce 14 phonemes per second. Even children show a remarkable aptitude in recognizing sounds as speech. Speech makes us unique creatures.
Language is understood in the research community [9] as well as in folk knowledge as central to being human. Because language is so central to being human, language processing has become synonymous with synthetic intelligence [10]. Understand how humans process verbal input, so the logic goes, and you will be able to build intelligent machines. For this reason a short overview of important concepts in synthetic speech is appropriate.
Synthetic speech research is often divided into two categories: Text to Speech and Automated Speech Recognition [11]. Text to speech (TTS) entails the creation of a sound pattern (voice) from a textual input (words). Automatic Speech Recognition (ASR), the inverse of TTS, entails the mapping of arbitrary voice input to printable text. While the field of ASR and the well known and often despised dictation systems have had for little real world success, TTS has made leaps and bounds in research as well as practice. TTS combines signal processing based acoustic representations of speech together with linguistic based analysis of text to create machinic utterances that sound like human voices. TTS systems are typically comprised of multiple components. A text analysis component defines and disambiguates the raw input. It finds sentence and paragraph breaks. It is also responsible for text normalization (mapping abbreviations and acronyms to full words). The output from the text analysis module is passed on to the phonetic analysis module. This module performs, amongst other things, the all important grapheme to phoneme (letter to sound) conversion. The output of this module, in turn, is passed on to the prosodic analysis module. It is charged with setting pitch duration and amplitude targets for each phoneme. Finally this output passes on to the speech synthesis module where the constructed string of symbols is rendered to an audible output reminiscent of a voice. TTS designers have experimented with various synthesis approaches for this last module. The most widely

used approaches today are concatenative synthesis and formant synthesis [11]. Here the concatenative approach is of particular interest. As opposed to the rule based formant method, concatenative synthesis is data centric. To construct an utterance, a concatenative TTS system would divide the input into segments, look for corresponding entries in a large database of recordings from a real human speaker (voice talent in the speech synthesis industry), and then concatenate (add serially) the individual parts to form the final output. This allows even sound sequences that have not been recorded per se to be rendered. The search, mapping and filtering steps included in concatenative systems are elaborate and deliver realistic machinic speech, particularly when perceived over low bandwidth media such as the telephone. Advanced concatenative systems include techniques of unit selection synthesis that automate the laborious task of (manually) finding correspondences, loosely speaking, between graphemes and phonemes. Unit selection synthesis is, in turn, heavily dependent on automated classification, most commonly implemented in the form of specifically designed neural networks the details of which are beyond the scope of this short overview.

Return of the Spoken Word

Join these technical advancements with the universally acknowledged significance of language and it becomes clear that TTS is of prime interest as a cultural phenomenon. Nothing less than a resurgence of oral traditions and a reassessment of the act of speaking can be expected in the wake of these new voice-centric systems. From the telegraph through punch cards to the keyboard and gaming console, computers have demanded people to meet them through clumsy haptic interfaces. TTS and ASR will spell out, literally, the end of the era of manually entered text input for machines. Furthermore, TTS and ASR redefine the quest for 'naturalness' in machines in ways other computer technologies do not. The consequences of this are far reaching, and this paper will only touch on some of them. But this much will be claimed: Speech technologies will allow for and require new definitions in our comfort zones with machines and with this they will create new hard (both in the computer science sense of intractable as well as in the cultural sense of multi-layered) problems in robot design.

The issue is not only academic. Many people share the feeling of discomfort when a gentle machine voice repeatedly cautions you to watch your step as you exit the escalator or experience anger when the menu of options offered by a kind robot voice of a service department you are calling does not in the least match your particular predicament. And even when robots do get it right, their tone of voice is often off. Statements have, in many languages, a simpler prosodic signature that questions, where prosody patterns vary widely as a function of the semantics of the question itself. This makes questions much more challenging to represent computationally than statements. Consequently, our intelligent speaking machines are better suited to issue commands than to ask questions.

But there is no turning back and synthetic speech will require us to think again about our own ways of expressing ourselves and how and when synthetic systems should mimic humans. The fact that machines can sound like humans does mean machines should use language in the same way people do. Beyond the flavor of utterances, synthetic speech begs the question of what machines could be saying to each other and what they should be saying to us. Cast as kind and patient, they have the capacity to say what we need to know, but also to insistently repeat what we have already heard or do not want to be confronted with. By default linked to databases and information systems, these übercorrect agents without a mother tongue have been delegated to the roles of clerks, instructors and supervisors. But they seem primed for more.

Accents and Immigrants

Speech acts not only reveal the intention of a speaker but also offer information on his or her origin. Many people who are born and raised in one culture and live in a different one, retain audible remnants of their past in their pronunciation patterns. Accented speech is particular speech because of the way its flavoring complicates the transmission of a message [12]. Prejudices alter the seemingly neutral transmission of content. Depending on the language competencies (and sympathy) of the listener, an accent can render an utterance charming, un-intelligible or even aggressive.

Digital signal processing can create arbitrary signals, most of which have no relationship to those human beings are capable of perceiving. With elaborate models of the human vocal tract's geometry, all sounds that humans can generate can be created artificially. Despite the universality of the technical infrastructure, TTS systems are usually designed for very specific applications along national fault lines with localized voice fonts and linguistically identifiable entities. Commercial vendors of TTS systems usually name these voice fonts according to their national linguistic origin, (but not by the individual voice talent who delivered the audio samples). For example, there are Sarahs for US English, Heathers for UK English and Günthers for German. The deliberate naming of these synthetic voices helps to enhance their believability; it conveys the comfortable feeling of a living person behind the digital audio utterance and assists in branding the product. The set of synthetic voices on the market represent a cleansed and controlled subset of popular human languages. It comes as no surprise that commercial TTS systems do not offer speech products with 'undesirable' features such as slurred speech or, say, a strong German accent.

Synthetic Heavy Accents: Make Language, part I

The perfect tone of the machine belies the fact that no human being really speaks without an accent, slight as it may be. We all come from somewhere and that somewhere flavors our lives and our voices. Only the machine can speak in a perfect tone that is location and history-free. To date, TTS and ASR researchers [13] have been interested in accented speech mostly for the difficulties it presents in intelligibility; i.e. when does an accent in a given language become a measurable hindrance in conveying a message - a recent important issue for telecommunication companies and call center operators.

In order to better understand the cultural fallout of synthetic speech, I am experimenting with synthetic heavy accents. In this context, I have crafted a German accented US English and a Mexican Spanish accented US English system with limited vocabulary [8] based on the SVOX speech engine [13, 14]. Several different methods allow one to craft accented speech. They range from combining a voice from one language with a linguistic model in a second language to recording an entire database from an accented speaker [15]. Even the simplest method of piping text from language A into a speech synthesizer constructed with a phoneme and grammar set of language B delivers useful results for some utterances. In order to generalize this language mixing, however, more elaborate methods are required. They include approaches gleaned from attempts to improve mis-transcriptions of ASR systems used to label grapheme to phoneme mappings [16], elaborate mixing of phoneme sets from two base languages, mixing of grammar requirements, modification of frequency, volume and word transition delays and various rules for exceptions. Both the German-English as well as the Spanish-English accents I have constructed combine all of these procedures.

The success of this process is dependent on the proximity of the languages in question. Indo-Germanic languages, for example, mix amongst each other with greater ease than with Slavic languages due to the similarity of base phones used for their articulation. The approach used here is brittle, however, and can not handle general purpose text or emotionally charged speech acts [17]. Also, the often awkward grammatical mappings that foreign speakers construct when they speak in a second language are not easily included in a formal grammar. The most general results would most likely come from building,

with no link to any base language, a completely new language model based on a particular accented speech, effectively treating it as a full fledged language. This would allow one then to construct any kind of accented utterance including those a human speaker would never consider making.

Synthetic Hissy Fits: Make Language part II

Can we learn something from mixing languages for synthetically accented speech that will help us imagine how we might mix human and synthetic beings? Imperfection, in all its rich variation, might be a a good learning ground for this. And from this we might better imagine what synthetic beings should actually have to say, once they are free from announcing flight schedules and processing concert ticket purchases. Consequently, we might consider investing more energy into telling machines about ourselves. Maybe our own language acquisition experiences can be ported to machines. When humans learn a new language they usually acquire an odd mix of bare essentials and examples of foul language. This is interesting as foul language circumvents the unknown new language and connects the speaker directly back to known territories [18]. While culturally specific in the boundary conditions that control its use, foul language links us more directly to our bodies than other forms of speech.

Fig. 1 Amy and Klara, 2006

Constructing a Potential for Synthetic Hissy Fits

In order to experiment with foul language in synthetic systems, I have built a set of nasty robotic agents housed in in cute pink boxes named Amy and Klara. Their ontologies are formed by and limited to reading and analyzing on-line trivia of life-style magazines such as Salon dot com. Creating ontologies that reflect common knowledge is an ongoing AI research agenda. Here, there is no claim to universality or completeness. This is a borrowed epistemology. The robots scan the website for news and cluster the results according to topics. This list becomes part of Amy and Klara's world. While the topics are meaningful for humans, they are but words to the robots. However, these words do receive, over time, significance by virtue of being repeated. Whichever topic is mentioned repeatedly receives more computational weight than topics found only once. Items that reach a critical threshold of numerically constructed significance become material for discussion. Amy and Klara share their statically weighted text summaries with each other via TTS and ASR. Since both robots scan the same

website, each robot expects the other to say what it already knows, and when this does not occur, dissent arises and they begin to call each other names. Each robot's ASR has a vocabulary of foul language, divided into increasingly aggressive scales of curse words. The robots were trained specifically to be able to respond to this kind of language that is filtered from ASR products. Additionally, the results from the speech recognizer as well as the physical transmission of utterances from speaker to microphone are error prone; miscommunication is unavoidable and with this, arguments practically guaranteed. The fact that Klara has a thick German accent only increases the likelihood for misunderstanding. Making the recognizer less selective creates more instances of false positive results (falsely recognizing a valid possible result) in which case the robots believe, as it were, to be offended by most all words they perceive.

There is no direct call to begin using foul language in the program of the robots. The arguments themselves are an emerging property of the above described configuration. Once one of the robots does emit a curse word, the other responds in kind if it recognizes the word. Each robot is equipped with a noise reducing microphone array and a small but dynamic equalizer-equipped speaker-amplifier system. The neural nets responsible for matching the perceived input to the foul language augmented linguistic corpus were trained with this hardware in place under low ambient noise conditions. Repeated use of a curse word from scale n leads to the selection of a curse word from scale n+1, provided the following word is recognized as a curse word within a given time frame (otherwise the aggression levels recede). Since recognition and utterance occur in quick succession, both a low level exchange (when recognition results are poor) as well as a heated escalating fight, if recognition results are positive, are possible. Furthermore, both robots are equipped with video cameras and able to see each other. Like the speech system, the vision system is geared for conflict. Amy and Klara both have a programmatic predisposition to be annoyed by the color pink, not knowing that they are pink themselves as each robot's own camera can see only white on the inside of its box. An adaptive histogram based hue detection algorithm [19] allows the robots to detect the other box's pink even under varying lighting conditions. The often idolized property of emergence is not limited to noble causes alone.

Of course the boundary conditions can be set to prevent arguments in which case the robots sit quietly next to each other. It is more interesting however, to set the conditions such that most episodes end in an exchange of expletives that leaves people watching the two robots really wondering about machine intelligence. A video documenting such an exchange of foul language is available on the web [8].

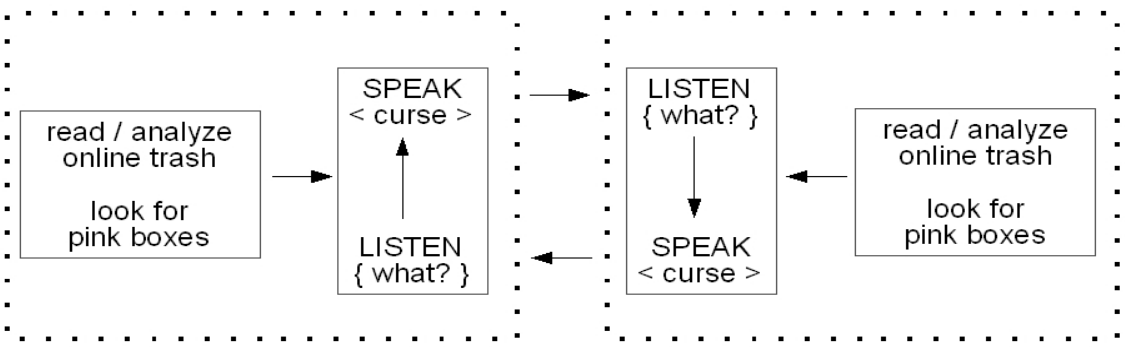

Fig. 2 constructing the potential for hissy fits in Amy and Klara

Learning from Amy and Klara

Amy and Klara and their provocative synthetic hissy fits warn us not to expect too much from

intelligent machines. They counter the rhetoric of the gentle intelligent machine with a critique of normative uses of synthetic speech and linguistic imperfection. Can we learn something about mixing robots and people by mixing languages in machines? If foul language is out of bounds for machines, then what about other taboos? Will we map all our taboos onto robots once they look, sound or smell like we do? Many linguistic taboos are derived from taboos in religion, sex and mental and physical ailments directly related to the physical constraints of being human. Since machines lack our bodily functions, the corresponding taboos really need not hold. We should expect some of our own taboos to be invalidated by machines. We should not be surprised if machines invent new curse words particular to the experience of being machine and having speech. Languages that are human in origin will be altered and amended by their use in machines in similar ways as popular culture alters and adds to the corpora of English language. There will be new figures of speech. Languages no longer in use by humans might be kept artificially alive in machines. And people lacking or having lost the capacity of speech might regain the skill in exchange with machines more patient than we are. Wolfgang von Kemplen, one of the first experimental researchers in synthetic speech, originally imagined his talking machine to be used for therapeutic purposes [20].

Fallout from advanced information processing technologies will make us continuously question our preconceptions of intelligence and challenge us to re-evaluate the ways in which we engage with machines. There is good reason to believe that human intelligence as we know it is dependent on the human body, and that language requires just such a body. Some computational linguistics researchers have attempted to prove this link in simulations of language evolution in embodied robots [21], and delivered astonishing results, provided one accepts initial conditions that allow for shared semantics [22] ab initio. However, if this is not given, the model fails. And so it might be helpful to reconsider the unconscious decisions we make based on an anthropocentric view of synthetic language.

Synthetic speech is a good example of the kind of dilemmas that perfect mimesis can generate. If nothing else, the experiments and observations described here might invite us to consider intelligence and cognition not particular to being human as worthy of attention. Computational devices have the capacity to 'be' in ways humans can not. The interior workings of machines and computational devices are so different from our own in material, construction, time scales and biological constraints. Being machine is not being human; rather it is a kind of foreign being that has no relationship to our own ways unless we force it to behave as such. What is lost in the mimetic approach of robot design is the opportunity to engage the otherness, as it were, of the machine. In this regard, engineers might be advised to check into the history of pictorial representation and its struggles over millenia with mimesis. From Zeuxis to Giotto, Leonardo and the Paragone of the Arts [23] through Caravaggio to the demise of the traditional art academy at the end of the 19th century, realism and mimesis have been potent and problematic principles of representation in Western civilization. Only in the wake of war and cataclysmic social upheaval did the arts find in abstraction and constructivism new pathways of non mimetic expression.

Synthetic speech research is a complex endeavor that demands rigorous attention to detail. Unfortunately, it is also another victim of the division of labor, as it were, that has established itself between the engineering sciences and the humanities and arts. This would be just another instance of a well known and often lamented disciplinary specialization if we did not have to repeatedly listen to the consequences on telephones and hear them in automobile navigation systems. How different might voice enabled machines sound and behave if they were informed by Wittgenstein's insight into meaning of words arising only from their use [24], or Rose's elaborately choreographed word games that begin with talk reminiscent of an academic presentation gone bad and end in a cacophony of utterances that sound like 'real' words but are nothing but babble [25]. Imagine if they knew about Blonk's powerful vocal tract, tongue and cheek skills that create sounds so odd they seem un-human and at times machinic [26], or the novelist Albahari, who surmised in recent work [27] the minimum number of words one actually needs to function. He counts five, provided one refrains from asking

questions.

Acknowledgments

This work is supported in part by a grant from the Humanities Institute at the University at Buffalo. Thanks to SVOX for making their text to speech engine available for strange experiments. Thanks also to FONIX SPEECH for their assistance in making their automated speech recognition engine handle Amy and Klara's hissy fits with relative ease.

Bibliography

[1] Pickering, A., "Cybernetics and the mangle: Ashby, Beer and Pask", Social studies of science, 2002, vol. 32, no3, pp. 413-437.
[2] Sorokin, P., "Social and Cultural Dynamics: A Study of Change in Major Systems of Art, Truth, Ethics, Law and Social Relationships", 1957/1970, Boston: Extending Horizons Books, Porter Sargent Publishers.
[3] CAE, "The Technology Of Uselessness," 1994, in: CTHEORY, http://www.ctheory.net/articles.aspx?id=59
[4] Kurzweil. R. "The Age of Intelligent Machines", MIT Press 1992
[5]. Fong, T., Nourbakhsh, I., and Dautenhahn, K., "A Survey of Socially Interactive Robots", Tech. Rep. CMU-RI-TR-02-29, Rob. Inst., CMU, 2002
[6] Simmons, R. Nakauchi, Y., "A Social Robot that Stands in Line", *IROS*, 2000.
[7] Ishiguro, H., "Android Science", Cognitive Science Society, 2005
[8] Böhlen, M., The Make Language Project, www.realtechsupport.org/new_works/ml.html
[9] Nass, C., Gong, L., "Speech interfaces from an evolutionary perspective", Communications of the ACM, Vol. 43, Num 9 (2000), P 36-43
[10] Christiansen, M. and Kirby, S., "Language Evolution: The Hardest Problem in Science?" In: Language Evolution: The States of the Art. Oxford University Press, 2003
[11] Schroeter, J. "Text to Speech Synthesis", in: Electrical Engineering Handbook, 3rd edition, chapter 16, p. 1-13, AT&T Laboratories, 2005
[12] Ikeno, A., Pellom, B., Cer, D., Thornton, A., Brenier, J., Jurafsky, D., Ward, W., Byrne, W., "Issues in Recognition of Spanish-Accented Spontaneous English", in ISCA & IEEE Workshop on Spontaneous Speech Processing and Recognition, Tokyo, Japan, April, 2003
[13] Pfister B., Romsdorder H., "Mixed lingual text analysis for polyglot TTS synthesis", *Eurospeech2003*
[14] Pfister, B., "Skript zur Vorlesung Sprachverarbeitung I+II Abteilung fuer Elektrotechnik", ETH Zuerich, 2005
[15] Tomokioyo L., Black A., Lenzo K., "Foreign Accents in Synthetic Speech: Development and Evaluation", *Interspeech2005*
[16] Kim Y., Sydral A., Conkie, A., "Pronunciation Lexicon Adaptation for TTS Voice Building", AT&T Labs, in: InterSpeech2004 – ICSLP, Korea 2004.
[17] Schroder, M., "Emotional speech synthesis: A review". In Proceedings of Eurospeech 2001,. volume 1, pages 561–564, Aalborg, Denmark, 2001.
[18] Hughes, G., "Swearing: A Social History of Foul Language, Oaths and Profanity in English". London: Penguin Books, 1998
[19] Bradski, G.R., Computer vision face tracking for use in a perceptual user interface. Intel Technol. J. 2(2), 1-15, 1998.
[20] von Kempelen, W., "Mechanismus der menschlichen Sprache nebst Beschreibung einer sprechenden Maschine", 1791 and http://www.ling.su.se/staff/hartmut/kemplne.htm
[21] Steels, L., Kaplan F., "Bootstrapping grounded word semantics", in: Briscoe, T., Linguistic evolution through language acquisition: formal and computational models, Cambridge University Press, 1999.
[22] Bickerton, D., "Language evolution: A brief guide for linguists", Lingua 117 (2007), p. 510-526.
[23] Steinitz, K., "Leonardo da Vinci's Trattato della Pittura: Treatise on Painting. A Bibliography of the Printed Editions 1651-1956". Library Research Monographs. Vol. 5. Copenhagen: Munksgaard, 1958.
[24] Wittgenstein, L., "Philosophische Untersuchungen," (1953), Suhrkamp 2003.
[25] Rose, P., "Pressures of the Text", video 17 minutes, 1983
[26] Blonk, J., "Vocalor" in: "Labior -Phonetic Etude #3", released on Staalplaat, 1998
[27] Albahari D., "Fünf Wörter", (2004), Eichborn Verlag, 2005

Glossary

synthetic speech	Speech created or processed by a computer
TTS	Text to Speech. The process of converting written words into audible speech.
ASR	Automated Speech Recognition. The process of mapping received audio input to text
phoneme	The smallest contrastive unit in the sound system of a language, independent of position in word or phrase
grapheme	The equivalent of the phoneme in written systems
prosody	Speech related information (intonation, pitch, duration, gestures) that is not contained in text itself. Prosody is often considered a parallel communication channel containing information that supplements or contrasts that of the primary channel.
concatenative synthesis	Synthesis based on the concatenation (or stringing together) of segments of recorded speech. This method generally produces the most natural-sounding synthesized speech but requires an often extensive database of recorded and tagged speech samples.
formant synthesis	Formant synthesis does not use human speech samples at runtime. Instead, the speech output is created using an acoustic model where parameters such as fundamental frequency (f0) and voicing are varied over time to create a waveform of artificial speech. Formant synthesizers usually produce more 'robotic sounding utterances but typically have a smaller footprint than concatenative systems as they lack a database of speech samples.
adaptive histogram hue detection	The process of dividing the complete color spectrum given in a given image into a a series of occurrence defined bins and using the results from this to set the boundary conditions for the histogram in a subsequent image such that a desired bin/color can be tracked over time.

Biographical Information

Marc Böhlen is Associate Professor in the Department of Media Study at the University of Buffalo and Visiting Professor at the AI-LAB at the University of Zürich. He has been contributing to diversity in machine culture since 1996.

Figure Captions

Figure 1 Amy and Klara's computers are synchronized (hence the large pink clock). This allows them to pop out of their respective boxes simultaneously. Each box is 10 x 10 by 10 inches. The robots are made of aluminum and plastics. Each robot is equipped with a microphone, a speaker and a camera

Figure 2 Schematic of the software architecture that allows hissy fits between Amy and Klara to occur.