“Do not attempt to light with match!”: Some thoughts on progress and research goals in Spoken Dialog Systems

Paul Heisterkamp

Research and Technology, Dialog Systems
DaimlerChrysler AG, Ulm, Germany
Paul.Heisterkamp@daimlerchrysler.com

Abstract

In view of the current market consolidation in the speech recognition industry, we ask some questions as to what constitutes the ideas underlying the ‘roadmap’ metaphor. These questions challenge the traditional faith in ever more complex and ‘natural’ systems as the ultimate goals and keys to full commercial success of Spoken Dialog Systems. As we strictly obey that faith, we consider those questions ‘jesuitic’ rather than ‘heretical’. Mainly, we ask: Have we (i.e. the scientific and industrial communities) been promising the right things to the right people? We leave the question open for discussion, and only cast glimpses at potential alternatives.

1. Introduction (sive capatio benevolentiae)

Ever since Spoken Dialog Systems (SDS) have become a reality, we, the researchers and engineers who are building these systems, have been working under a number of assumptions pointing towards a human-like recognition, understanding and dialog behavior of ‘The Ultimate Speech System’. Of course, we knew and know all along that under the limiting conditions of finite hardware, we have to make do with limited domain systems. Yet we perceive them as being steps towards the ultimate system, coming with increased vocabularies, robustness, dialog capabilities, better speech synthesis etc. in a constant movement towards that goal – and, apparently, for a long time we have turned out systems with the respective improvements. Consequently, the industry growing up around SDS and their components has promised their customers ever more natural dialogs. The assumption here is that ‘natural’ equals ‘easy to use’, which in turn assumes that this kind of ease makes for a higher consumer acceptance and therefore for better business.

In the last year, we have seen a considerable concentration in the speech industry. At least in part, this concentration is due to the fact that the losses speech companies have been accumulating (with rare exceptions) are no longer covered by an abundance of venture capital. A number of SDS in commercial use are very well designed and very successful. Still, we do not see the wide breakthrough and public acceptance we have been anticipating, nor the respective economic success. In view of the crisis, companies are now even lowering the prices for the deployment of complete systems by letting their customers (deployers) take over part of the development by making available dialog description systems like VoiceXML or SALT for them such that people with little to no experience in the design and implementation of SDS now can take over these tasks. The obvious threat is an abundance of ‘cheap’ systems.

In this paper, we raise a few questions. First, we want to start a discussion about the validity of the claim that naturalness and ease of use are the equivalent and that the former is a prerequisite of the latter. We argue that this is not necessarily the case and give some anecdotal and historical evidence. We choose as an example the often-heard demand: “Speaking to a computer should make things as easy as switching on the light”. We continue to ask as to what might be an approach to *interim* systems, i.e. such that, while research continues on the road to ‘the Ultimate Speech System’, the SDS actually deployed are not overloaded with the – perhaps too ambitious – demand to be partial realizations of the great goal. We lay out some points we see as critical in research and education in the next years for the realization of interim systems. And we ask the question as to how the community could communicate a change of perspective (should it occur) to customers and funding agencies.

In terms of the ‘roadmap’ metaphor that underlies this workshop, what we see in the autobahn picture is a projection not so much from today’s systems into the future, but rather a perspective from the future ‘Ultimate Speech System’ and a partitioning of from that point back in time towards the present. The terrain that lies between this future and now is, as yet, unprospected. From where we stand we can see, perhaps, some hills and mountains of problems arising from the ground and lay our track such as to overcome or go around them, and in this respect, it is definitely justified use the roadmap metaphor. What this paper tries to do is to raise the awareness that there may be trenches, ditches or boggy spots even on the path laid out for the autobahn, invisible from our low vantage point, where straightforward road construction does not help, where we may need auxiliary structures like bridges or tunnels and where, to exhaust the analogy, an off-road vehicle is more successful that the sleek sports car that will finally travel the autobahn.

Disclaimer: This is a discussion paper for a workshop. The thoughts laid out here do not make any claim to completeness or even consistency. The aim is to take a step back from ongoing work and perhaps start a discussion in the community: The author professes to believe in ever more complex and natural SDS to come and in investing every available resource to that end. (And, dear reader, forgive us for showing off a little).

2. The Ultimate Speech System?

Why do we want to build technical speech systems? Because we want to deliver better and cheaper services to people. Speech systems with their ‘natural’ interaction promise to make available the full power of our computerized world to everybody. You utter a whish, and it is granted. Some people (e.g. Pakucs [1]) have introduced the metaphor of a 'butler' for this, others call them 'fully conversational' or 'Say Anything Anytime'. The idea is to have the full human understanding capability combined with an expertise on all available application systems and their potential interactions, plus a very good idea of what the speaker (or master) really wants. We have come quite a long way to realize parts of such system, so far, in fact, that now we begin to see practical limitations that may not be surmountable with technology alone. The better the technology works, the more important the human user becomes as a part of the overall system.

2.1. People

In deployed systems, both in telephony and in embedded Command&Control systems, a considerable portion of the 'errors' (in the sense of the system not doing or responding as the user anticipated) is due to 'Patron errors': People speaking not to the system, people not saying what they want, people not providing the information requested by the system. From practical experience, we estimate that any SDS can only have about 97% task completion rate if there are new users involved, regardless of the system's recognition, understanding, dialog etc. capabilities. Human operators can only perform better because misunderstanding and inherent error correction are part of human communication, and often go unnoticed. Speech production is an evolutionary much older and less consciously controlled process than writing, and thus more prone to errors, a fact many people in linguistics tend to forget ("Wherever you read 'ill-formed input', replace it with 'naturally spoken utterance'." Nick Campbell [2]). To illustrate the inherent human repair capacity, let us do a little experiment (cf. [5]). Read the following sentence and give a quick answer:

(1) "How many animals of every kind did Moses take into his Ark?"

Wait! Got it? The answer 'two' is correct. Now try to repeat the sentence (1) in your mind (don't look!). Repeated it? Good! Now answer this question:

(2) "Who built the Ark?"

The answer is, of course: Noah. Even if this did not work for you (E.g., this only works if you have a judeo-christian background), most people give the correct answer 'two' to question (1), repeat the sentence including the 'Moses' part, and only in re-thinking it can answer question (2) correctly. (1) is, in our view, an example of a very distorted signal that is implicitly corrected in semantic processing. Without going into details about constructivist views of perception and understanding, suffice it to say here that we have strong doubts as to whether this kind of semantic overriding, through ontology-based plausibilities or other mechanisms, will be computationally available in a wide scale in the next five to ten years. The question is, even, if this kind of processing does not in its turn introduce some unreliability we would want to avoid in a technical system. Thus, there is a part of human understanding capabilities that may, at least for a long time to come, be unattainable by technical systems. People don't always mean what they say. Now, do they say what they mean?

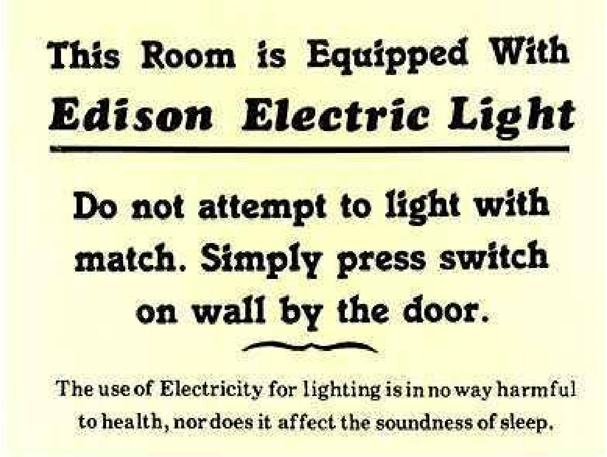
2.2. Applications

One of the intriguing aspects of SDS is that they are on the one hand technical systems and on the other 'natural' communicators in the sense that they take over the role of a dialog participant, a role that normally can only be taken by humans. We have argued earlier that SDS attempt to bridge the gap between technical systems that require clear and unambiguous commands, and humans who think in terms of problems to be solved. Now, for some problems, one can safely assume everybody in the western culture has some understanding of what possible solutions are. However, these solutions have a historical and cultural background. We will come back to this later on. Where SDS replace services that people know or offer an alternative and perhaps more sophisticated access to 'known' solutions, today's systems are quite successful, provided the developers invest sufficiently in a good model of the application and the dialog (cf., e.g., [3] for a 'one-shot' system).

A problem arises with new or very complex services and solutions. How can people address a system may or may not be able to solve the particular problem they have, or might even be able to provide services they never heard of? Designers try to overcome this difficulty by designing help systems or introductory tours or self-explanations etc., but all of these approaches suffer from the same deficiencies we all know from dealing with non-speech computers: These introductions are tedious, they eat people's time and often overstretch their attention span. Especially in the case of telephony systems, these introductions may even cost people's money. And, more important in our context, these help (sub-)systems are in themselves a way to educate the users, something that is at least slightly askew with the assumption that naturalness equals ease of use. The question arises here whether this assumption is valid.

3. 'Natural' interaction

A person comes into a dark room. The person wants to be able to see. For us today, we think it natural to tap with our hands in the region next to the door to find the light switch. We find a switch, we press it, the light goes on and our problem is solved. Around the end of the 19th century, this behavior was by no means 'natural', as witnessed by the following sign:



This Room is Equipped With

Edison Electric Light

Do not attempt to light with match. Simply press switch on wall by the door.

The use of Electricity for lighting is in no way harmful to health, nor does it affect the soundness of sleep.

Figure 1: Sign reportedly found in New York City hotel rooms at the end of the 19t^{h} century (from [4]).

In terms of human-machine dialog, the process of ‘switching on the light’ in our modern sense implies a number of assumptions that have developed over time, and, just as a pidgin contact language becomes a creole language in its own right when it has native speakers, for those who have grown up with this process as the normal way to solve the problem of the dark room, this process is the standard and may well be called natural without quotation marks: Every child can do it, and you do not have to understand any of the technology or the details involved. Some of the assumptions in the dialog between the system (light/switch) and the user are:

- A room has an electric light
- The light is operated by a switch
- The switch is in the room or outside near the door
- A switch in the room near the door ‘belongs’, or is connected to, the light in that room
- The light is working (bulb, power, cables etc.)

The ‘switching on the light’ process is a dialog: if the designer of the system obeyed the conversation conventions, the presence of a switch in one of the habitual locations can be seen as a communicative act (system-driven, so to speak), offering a communication:

(3) “If you have a darkness-problem, interact with me, and I’ll solve it”

The communicative act of pressing the switch equals an acceptance of this offer. If the light does not come on (the confirmation act), you can observe what this ‘broken promise’ in communication does to people: A very normal first reaction is to press the switch again and again, and with more energy, i.e. to repeat the communication act with more emphasis. Only later do they start to question the dialog assumptions, e.g. look for another switch or check (using other switches) whether power is available (and finally insert that hotel key card in that box in the dark), etc.

Switching on the light is easy – a very easy interface to an enormous and complex energy supply system, but, just as this system, the dialog is not natural! At least in the beginning, it had to be learned, i.e. the deployers of the system (Edison!) had to teach their potential customers (guests in the hotel could have b{r}ought a candle!) how to access their services. The advantages of electric light over the other means available at that time (notably gas) convinced people fast, so that for us, the native ‘electric light creole’ speakers, all this seems completely natural. Of course, there are dialects of this language: For a European, finding and operating the turning switch of an American bedside table lamp, near the bulb and often obscured by the lamp shade, or for an American finding the switch of a European bedside table lamp attached to the cable and slipped behind the bedside table can be a several minutes exercise. The existence of these dialects demonstrates that even here, in a relatively short period of time and keeping the same conversation maxims, the Saussurean arbitrarity of the sign has given way to conventions.

Another interface to a complex system originated around the same time at the end of the 19th century. It was (and is) by no means easy, yet, through these conventions, is has remained virtually unchanged despite the fact that the technical limitations that lead to its installation are not a real issue any more, and despite the fact, also, that there have been numerous efforts to reform: The typewriter keyboard layout, commonly known in English-writing countries as ‘qwerty’.

Most people, even in computer industry, believe that this layout is the way it is because it enables people to write the fastest way possible. Now, this is not the case. In fact, the ‘qwerty’ layout originally is a compromise between the desire (and the promise) to write as fast as possible, and the timing problem mechanical lever-type typewriters have. The type-lever, having struck the ribbon and impressed its character on the paper, needs a certain time to fall back to its original position. If another type-lever is raised across the first type-lever’s fall-back trajectory, the type-levers get entangled. The typist then has to stop and disentangle them, there may even be serious damage to the machine. The ‘qwerty’ layout reduces this problem by placing often-used two-letter combinations (e.g. ‘er’, ‘sh’ for English) such that the physiognomy of muscles, sinews and bones of both hands require a certain amount of time before striking the respective key sends a potentially obstructive type-lever upwards ([6]).

The original reasons for the particular ‘qwerty’ layout are long since obsolete, but neither reasons of speeding up type writing nor making this keyboard easier to learn or to memorize were able to develop sufficient attraction for people to sacrifice convention (call it upward or downward compatibility, if you wish) for their sake. Even most people who otherwise modify their computer keyboard at will (e.g. emacs wizards), never change the original layout.

4. A question to the author

(Admittedly, the question is rhetorical; but what would you expect?):

Now, what has all of this to do with the future of Spoken Dialog Systems and the road that leads there?

In the examples above we first wanted to raise the awareness to the fact that some things that seem natural to us, at least in interfacing with complex technical systems, are, as it were, conventions. These conventions were not natural at the time of their origin. They were, in their beginning, the kind of auxiliary constructions that helped their technologies being successful in spite of these technologies’ limitations and although people couldn’t interact ‘naturally’ with typewriters or electric light. They were not necessarily simple. Yet, they were successful. They have survived over a hundred years, They have survived regardless of whether their original rationale is still valid.

They can’t be all *that* bad, these conventions, can they?

Second, the example should illustrate that conventions may be useful, if not exactly needed, at the introduction of new technologies. While we continue to think that every effort to reach the ‘Ultimate Speech System’, we ask: Could we get better intermediate systems (and better business), if we, as the community of researchers and industry people, would divert some of that effort to create and publicise a set of conventions that helps people get along with the not-so-perfect systems we will turn out during the next years?

5. Establish Conventions!

A convention differs from a standard (and be it an industry standard) in that there is no need to have a committee agree on things, or people coming together who have certain intellectual property rights to protect. It does not even have to be outspoken!. A convention is ‘open’ in the sense that everybody can choose to conform to it or not, but it is established only if many (and major) players keep to it. As there are fewer (and bigger) players on the market now, this point in time may be the right opportunity to start this discussion. We are neither in a position nor able to give a recipe how to establish conventions speech research and industry. We just mention a few items for discussion.

One approach to establish conventions is to transfer existing conventions from another field. The ‘Speech Graffiti’ group at CMU (cf. [7]) does this. For form-filling dialogs, very common in database queries etc, the idea is that people can make good guesses at the names of columns of, e.g., relational databases. Now, a ‘generic’ speech interface to such a databases takes the names of columns as attributes and together with the names of values to incrementally build a query. Of course, some lexical synonyms are allowed.

The intriguing thing about ‘Speech Graffiti’ is that it makes use of ideas about the application layout that are common knowledge among the ‘computer literate’. However, we see this is also the drawback. It does replicate line-typing SQL-style database requests. If you’re not familiar with those, bad luck! Still, in our view, this type of generic interface could help establish conventions in the ‘help’ sector: People can ask: ‘What can I talk about?’ and get back a list of column names.

There is a problem with transferring conventions and replication interaction styles: What is good practice for one may be horrible for the other. SDS have suffered a lot from people replicating DTMF (touch-tone) interfaces in ‘speech’. Worse may yet come: with the wide distribution of VoiceXML and/or SALT dialog tools, anybody who can speak can build a bad dialog. A fool with a tool is still a fool. Some SDS we have seen recently let us take serious the warning of Jim Larson, co-chairperson of the W3C web browser initiative, who fears that an abundance of bad systems may eventually lead to a drop in market and funding for the speech community: another ‘Speech Winter’ (e.g. [8]).

As we can’t make these dialog tools vanish, what, then, can we do? The current reduction of the number of different speech technology suppliers should make it possible for their sales people to honestly tell their customers: “Hey, it’s neither easy nor cheap to get a good system. You need decent modeling, careful prompt design and a number of other things to make speech a commercial success, not just recognition rates. If you want a good deal, invest in the dialog design as well!”

Of course, the speech community would have to match this be educating their students not only in engineering, but to make sure that speech engineers also have some understanding of dialog design, and the importance of seeing the overall system, including the human user. Take s step back from looking at single components which forever need improvement, and see what the benefit for the overall system is. Academic research and teaching might give a little more attention to initiatives like the DISC project [9], of which little is known, as job interviews show.

So now here’s your discussion starting kit:

- Don’t make it easy to build (bad) systems.
- Invest in teaching design!
- Invest in educating customers!
- Don’t promise ‘ease’ and success through better recognizers
- Admit that efforts must be made in modeling users and applications and re-design during operation!.

6. Concluvision

By investing in establishing SDS conventions, we may bind resources such that we get the ultimate speech system a little later, but we have the chance that more people get the idea that Spoken Dialog Systems can deliver, easier, better, faster, more reliable services:

In the movie ‘Small Soldiers’ (cf. [9]), the adolescent hero tries to place an urgent request for help on a product in a company’s call center. The call center agent firmly denies him being able to make such a request, as the respective product is not for sale yet. Finally, the youth yells:

“Hey, listen, isn’t there a machine that I could talk to?”

7. References

[1] Pakucs, Botond (2003): “SesaME: A Framework for Personalised and Adaptive Speech Interfaces”. In: Proc. EACL-03 Workshop on Dialogue Systems, Budapest, Hungary.

[2] Campbell, Nick (1996): Oral comment on a paper at the 1st ESCA Workshop on Interactive Dialogue Systems (IDS ’96), Vigsø, Denmark.

[3] Peckham, Jeremy (2002): Can automation be an improvement on human agents? In: Proc. LangTech, Berlin, Germany.

[4] Mijksenaar, Paul; Westendorp, Piet (1999): Open Here – The Art of Instructional Design. New York. (Remark: We have some doubts as to the sign being genuine. To the best of our knowledge, electric switches at that time were flipped or turned, not pressed. This does not, however, reduce the argumentative or entertaining power of the book; nor, hopefully, the power of the argument in the text).

[5] Sorry for the breach of order; otherwise, the figure caption for [4] woudn’t have fit on that page. Now for the reference. There is no reference. We have been using this conjurer’s trick at least since 1989. It has appeared somewhere in psycholinguistic literature, and it works well, but the source is long forgotten.

[6] No real reference for this could be found short notice; most of the text quoted can be found (in German) under the URL http://www.fragenohneantwort.de/sonstiges.htm

[7] Rosenfeld, Roni (2000ff.): USI or ‘Speech Graffitti’ (-tt-sic!), e.g. http://www-2.cs.cmu.edu/~usi/ (The description in the text should be taken with a grain of salt).

[8] Larson, Jim (2003): Introductory talk to current W3C voice browser and VoiceXML activities, at ‘Voice Enabled Services’ business conference, London.

[9] Spoken Language Dialogue Systems and Components: Best practice in development and evaluation, EU project 1997-2000, URL: www.disc2.dk

[10] Small Soldiers (USA 1998), Director: Joe Dante.