SPEECH DIALOGUE SYSTEMS IN THE TELECOM NETWORK

V. Steinbiss

Philips Speech Processing
Kackertstr. 10, 52072 Aachen, Germany
volker.steinbiss@philips.com

ABSTRACT

There is good reason why the most impressive success stories of speech recognition are reported from systems installed in the telephone network: The technology fills the gap between proficient but expensive human operators on the one hand and the ergonomic hell of DTMF-based systems on the other. At the same time, there are enormous business opportunities that can be realized with speech dialogue systems. Meanwhile, an increasing number of interesting system deployments both gives us an idea on the validity of technology and business cases and indicates how voice-operated services in the telecom network (or, well, in the Internet) might look like in the future. We will take a closer look into the probably most interesting applications, namely voice portals, directory assistance, and customer services.

Looking into the nearer future, our world will change even more dramatically by speech recognition enabling Internet access without a graphical browser, or on a (mobile) device with limited input capabilities. This will accelerate convergence between the telephone network and the Internet and bring a variety of Internet-like services to mobile and phone users.

1. Speech Recognition in General

Speech recognition systems, and their communicative extension speech dialog systems, will change our world, just as other technologies have done before – e.g. the telephone, the personal computer, the mobile telephone, and the Internet (still ongoing). As the future is always difficult to predict, let us approach the subject from several angles, including some basic pro's and con's of the technology (chapter 1), the (end) user perspective on its added value (2), some more technology (3), some example applications (4), and a wish list to the academic world as seen from a technical person who would like to enable people to interact with technology in a most natural way.

Apart from the creation of text – an important application of speech recognition, especially in the professional areas medical, legal, and insurance – the purpose of speech recognition is to help communicate with computers in the broadest sense, including electronic devices like mobile phones, televisions, etc. This human – machine interaction takes advantage of the fact that human beings are very good in talking, and especially better than using a keyboard or buttons to control menus. From the ergonomics perspective, speech has the advantage of being natural, requiring no or only little mental transformation. E. g., to phone Volker Steinbiss, a command "phone Volker Steinbiss" requires less mental transfer work and so is certainly more pleasant than pressing the button sequence [Names] - [Name search] - [7PQRS] - [8TUV] - [down] - [down] . This advantage holds in many interactions, where the number of potential choices is large – speech is a very broad-band input channel. It need to be stressed that it is *not* expected that speech will *replace all* the other input modalities; however, it will replace a few and coexist with many.

As speech enables hands-free and eyes-free interaction, it is well suited to support mobile usage scenarios, including Internet access from anywhere. There will be situations where the user has the choice between different input modalities, and – like today with mouse and keyboard – there will be different preferences depending on the application and user type.

2. People's Needs

Two models of spoken human machine interaction can be a starting point when thinking about future usage: first, the button or menu oriented way of interacting with devices, typically with some graphical feedback, and sometimes with a voice feedback as in DTMF based interactive voice response (IVR) systems, and, second, human-human communication. There is indication that pure button replacement does not deliver acceptable user interfaces, as the power of speech is not fully exploited and at the same time constrained by the old interface design. A horrible and completely unacceptable way of using speech is demonstrated by systems of the *"press or say 2"* kind: While the IVR system kind of mapping *"to call the marketing department, press 2"* can be excused as one cannot do much better with just a DTMF tone generator, there is no excuse for a system interface where a caller has to *say numbers* to retrieve *menu items*. Systems with well designed dialogs where people say meaningful items are much better, even if they use rather simple recognition technology.

A problem that goes with strongly constrained systems and directed dialogs is that people often have difficulties to understand what the system is capable to understand, and what it isn't capable of. If the system status changes during the dialog, this is difficult to communicate to the caller. Ideally, it should be avoided to force callers into the framework of a technical system – they should not bother to understand the internals or the restrictions but rather express themselves in a natural way – having in mind human-human conversation, still knowing that they talk to a machine. Callers should be given guidance or help when they want it, but they should have the power to be the leading part in that conversation. Machines should serve people, and people should be in control.

3. True Natural Dialog Systems

The straight-forward approach to dialogs (at least straight-forward for technical people), the directed dialog, breaks the task into little parts and executes them like a computer program, step by step. The system might ask for a credit card number, then – after successful input – for the expiration date. Before completion of one of these steps, it might be necessary to go through several dialog turns, e.g. in order to confirm the input.

This approach is good enough for tasks that somehow impose a natural sequence of steps which is plausible to the caller. In many situations, however, it is possible to guide the caller but only in a way which feels unnatural – people feel constrained by technology. VCR programming has become a well known example for this. So, how can we make dialogs more natural and human-like?

To achieve this, one has to re-think how a dialog should be executed. In the mixed-initiative dialog systems of Philips Speech Processing ([1]), there is no explicit coding of single steps that have to be taken one after another. Instead, the system is equipped with a search engine that in each dialog situation takes the next step to fulfil the task – typically filling slots of information. In the prototypical example of a train inquiry system – the first mixed-initiative spoken dialog system that existed – the dialog engine has the goal to get four pieces of information: the departure and the arrival station, the date and either departure or arrival time. Heuristic rules trigger the next dialog turn of the system. To give an obvious example, the system asks for the next piece of information which is still missing. But there is more to it: As the input is speech, and recognition errors might occur, the system follows a verification strategy. Before seeking for more information, it verifies that the recognized information item was really what the user said. Depending on the application, this might be explicit ("Did you say 'Berlin'?") or implicit ("At what time do you want to arrive in Berlin?"). Keeping track of the status of information items, such as recognized, verified, or other information of previous dialog turns (such as that a value had been negated, or a list of hypothesis with probabilities), and processing this information accordingly, has a strong impact on how the system is being perceived by a caller.

The two examples will illustrate this.

If a caller calls at 3 p.m. that he/she wants to take the train at 6 o'clock today, the system first computes the value of this time concept, which consists of two times – {6 a.m., 6 p.m.} – and then combines these values with its knowledge base, concluding that the time which lies in the past can be removed from the hypothesis list. This system asking "6 p.m., right?" is perceived superior to a system asking an obviously silly question ("6 a.m. or p.m.?"). Note that some of these design decisions might be preferences rather than hard decisions – e.g. a caller might still want to know something that happened in the past. In this case, the statistical framework of the recognition and understanding components makes is powerful enough to assign to these unlikely interpretations probabilities greater than zero, such that the caller can insist on them while the default system behavior would still be different.

In the second example, consider an automated attendant which puts callers through to people in a company. While some queries would immediately lead to a unique telephone database entry ("Connect me to Mr. Vohringer, please"), others would create a list of options that has to be combined with information from successive dialogue turns:

CALLER: "I would like to talk to Mr. Smith, please."

SYSTEM: "Which Mr. Smith would you like to talk to?"

CALLER: "He is in the marketing department."

SYSTEM: "I connect you to John Smith – pls. hold on."

From the caller's perspective, there is a lot of benefit using mixed initiative systems with decent understanding capability. Callers might start as novices, using a system for the first time with little understanding of the offered functionality, then pick up fast and mutate into power users who want to get the information fast and effective. Mixed initiative systems offer support and guidance to the novice and extremely fast execution to the expert user.

While guided dialogs might do in many situations, their significant drawback is that they put people into the schematics of a dialog flow that has been designed by someone else, just as in VCR programming. We should not constrain people by technology if not needed, and the need is not there any more – the technology is proven by many installations [1], the principles have been published [2], and the tool-sets for developers are available [1].

4. Applications in the Telecom Network

Today, telecom networks face increasingly fierce competition. Market changes in the United States and in Europe have created a new playing field with a host of aggressive players. Customer access to a variety of new and improved services is paramount, but so is cost reduction.

How can a company offer user-friendly services while keeping prices competitive? High-end dialog systems are addressing these needs by either fully or partially automating these services. High-level speech recognition technology enables telecom operators to reduce costs, improve customer service and attract new customers. But even more important than just cost reduction, the new technology offers a range of applications that have not been possible before with either operator services, simple DMTF based IVR systems, or stationary GUI-based Internet access.

The examples for services offered with speech recognition are practically limitless. Some of them are: Information services, such as weather information, travel information, news services; ordering services (taxis, movies, flowers, etc.), yellow pages (restaurant reservations) and white pages.

Let us look into a few interesting applications that give an idea of the “new world” of telecom services.

4.1 DIRECTORY ASSISTANCE

Directory Assistance, both white and yellow pages, is an important and highly frequented service. As in other areas, the provided functionality is for the lion’s share of the phone calls just a simple database (or phonebook) lookup that has to be performed by human operators due to the fact that phonebook or Internet access are not at hand, or that people just prefer to call. It is commercially unreasonable to perform this kind of task with operators once there is an alternative, which is now (1Q2000) coming with the first speech-recognition based directory assistance systems [1]. Quite comparable to customer contact via the Internet, the availability of automated directory assistance will result in dramatic savings of cost per call. It is well understood that, like in many other applications, the operator might want to handle a fraction of the calls via operators. The advantages of an automated system might be given to the end customers in form of reduced cost or reduced waiting time.

Technically speaking, speech recognition is just maturing to handle this task which is characterized by a very large vocabulary and low redundancy. The inherent ambiguities due to recognition uncertainties (*Schmidt* versus *Schmitz*), homophones (*Schmidt* versus *Schmitt*) and other ambiguities (names just differing in first name or even street) can be successfully hidden to the caller by combining all knowledge collected in the dialog and matching it accordingly with the database. Again, [2] is referred to for further reading and references on the technical issues behind these systems.

4.2 CUSTOMER SERVICES

For customer services, the same reasoning holds: Some telephone based services should be performed by operators, but many others should be based on speech recognition and dialog technology. An impressive customer services installation at the leading Italian mobile operator Omnitel [3], called Omnitel 190, demonstrates the viability of the approach. Omnitel has just surpassed the 10 million figure of subscribers. Their automatic system, based on Philips’ SpeechMania^{TM} platform, handles some 200,000 calls per day, of which most are automatically processed.

As described in the technical section, users are free to either get guidance or go directly to their area of interest.

4.3 VOICE PORTALS

The difference between a suite of telephone services and a voice portal offering access to services seems to be small: Customers just have to remember the single one number of the voice portal as compared to several numbers otherwise. And, when calling the voice portal, customers do not have to navigate through a lengthy menu but just say which service they would like to access. This “little difference” is expected to have a huge impact as it eases service access – it basically transports the concept of an Internet portal to the telephone world. As a difference though, the first systems will just offer the suite of services of one service provider, or service providers that they cooperate with – as a reference, see the world’s first voice portal, Omnitel 2000 [3].

An interesting upcoming description language for directed dialog, VoiceXML [4], which already contains some mixed-initiative dialog elements, has in its first version 1.0 already envisaged to load dialog description scripts dynamically (depending on the dialog flow). This will enable the caller to jump through Internet based voice applications as one jumps from web site to web sites these days. – A remark: I do not differentiate here between voice operated telecom services and voice operated Internet access without graphical user interface, as there is no relevant difference.

4.4 MOBILE INTERNET ACCESS

An interesting range of applications is personalized to one user, e.g. personal call assistant services, such as voice dialing, voice and fax mail, unified messaging and personal organizers. There is a smooth transition from the voice portal to more these personalized services and to access to offers that currently reside in the Internet, in a form that is typically accessed via a graphical browser. While many people tend to associate the Internet with graphical user interfaces, there is reason to believe that the Internet's other (and I believe more important) characteristics digitization, computing and hyperlinks will lead to increased mobile access to services, data etc. via voice. Again, this might be called mobile Internet access or voice operated networked services via the telecom network. The future will show what the typical mix of modalities will be. While driving a car, speech-in and speech-out seems to be the method of choice; while in a meeting, speech-in is tabu; and in many other situations, the preferred mode could be speech-in and graphics-out – for each channel, use the fastest carrier. Multi-modal input with a strong speech component and multi-media output with a strong graphics/video component are most likely to occur but not far developed yet.

There is little doubt that the number of automated speech enabled services will strongly increase in the coming years. One of the key success factors will be the ease in which the companies who provide theses services, e.g. the telecom operators but also many others, can set up and modify services. This depends to some extent on the recognition and understanding performance of these systems, the tool-sets, and the availability of language resources. As the market is still young, there is some lack of standards and even more some lack of experience in building successful applications. At this point in time, the bottleneck is no longer in technology but in picking up the technology and making it a mainstream business. I expect the main drivers here to be the aforementioned applications directory assistance, customer services, voice portals, and mobile Internet access.

5. More Technology Is Needed

Many public funding programs seem to reflect the thinking that the speech recognition problem is almost solved and will be done somehow anyway. This is only right to the extent that powerful working systems exist and that the speech recognition industry puts considerable resources into research and development. However, both academia and industry agree upon that much progress is still to be made, as recognition systems these days typically function quite well in the situation that they have been designed for but tend to drop significantly in performance once training–test mismatches become strong. There is still much work to be done, e.g. in the areas of robustness against various forms of mismatch, in adaptation, understanding etc. Commercially relevant and also a technically challenging is how to set up an application with very low effort – to describe dialogs simple and effectively, to avoid too extensive data collection, etc.

Looking into technologies that are by some means adjacent to speech recognition, understanding, and dialogs, there is still much more to be done than plugging the pieces together.

Concerning human human interaction, the guiding task here is to arrive at the translation of spoken language. The BMBF funded project Verbmobil [5] has contributed valuable research here and is a source for further reading, as well as the C-Star project [6]. There is still much progress to be achieved concerning machine translation.

Human machine interaction is also worth being a focus of research. Speech recognition and understanding has to be improved using prosody and detecting mood in speech. The interaction via the phone will strongly benefit from enhanced user modeling and personalization issues. At least for the home environment, we have to add other input modalities like eye tracking, lip reading, gesture and face recognition.

Much better technology is needed so that finally people can forget about technology.

6. REFERENCES

[1] http://www.speech.philips.com/
[2] Souvignier, Kellner, Rueber, Schramm, Seide: "The thoughtful elephant", in *IEEE Trans Speech and Audio Processing*, Vol. 8 (1), Jan. 2000, pp. 51-62.
[3] http://www.omnitel.it/
[4] http://www.voicexml.org/
[5] http://verbmobil.dfki.de/
[6] http://www.c-star.org/