City University London

OFCOM

CAPTIONED TELEPHONY

FINAL REPORT

EXTENSION OF 2006 RESEARCH REPORT – "FEASIBILITY OF ADDITIONAL TELEPHONE RELAY SERVICES"

CITY UNIVERSITY RESEARCH TEAM

City University London

CAPTIONED TELEPHONY

CONTENTS

INTRODUCTION

With "Voice Carry Over" telephone service, a person who is hard–of–hearing or deaf can speak with his or her own voice, if preferred, and can receive responses from a hearing person in the conversation by means of typed text (relayed by an "operator" or "communications assistant"). This is one option provided by the UK text relay service RNID Typetalk, and relay services in other countries.

Captioned telephone service is an enhanced Voice Carry Over (VCO) service that allows a user to both listen to what the other party is saying and near-simultaneously to read captions of what the other party is saying. In this way, a typical user of this service, who has the ability to speak and has some residual hearing, can both listen to what is said over the telephone and read captions for clarification. A third party in the telephone conversation is a communications assistant (at a location remote from both conversants) who generates the captions. In current commercial services these captions are generated by the communications assistant re-voicing one side of the conversation into a computer running voice recognition software to generate the text.

The service's captions may be displayed on a special telephone, which is equipped with a screen, or (recently available) on a personal computer. A key aspect of captioned telephone is that it maintains nearly the same level of spontaneity as a typical voice-to-voice telephone call. The dialogue is closer to synchronous communication than the asynchronous methods of traditional relay. A captioned telephone user can speak directly to another party with his or her own voice, listen to the actual voice and inflections of the other party, and read the text of the conversation to support and clarify what is heard and understood.

In the UK, there is a single captioned telephone service provider using the PSTN – Teletec Ltd., Cranfield, Bedfordshire. This company has been offering the CapTel service, only usable with their special phone equipment incorporating Teletec's proprietary technology. The CapTel service is no longer offered to new customers (but remains available indefinitely to existing customers). Teletec has recently made available a WebCapTel service which provides captions via an Internet-connected device (whether computer, mobile phone or PDA [Personal Digital Assistant]). The voice parts of the conversation may be carried by ordinary PSTN-connected telephone equipment and also works with VoIP calls. The WebCapTel service also uses a re-configured variation of the company's proprietary technology which is now contained in the software and communications protocols of the service, instead of in the phone hardware (as for CapTel).

In the US, captioned telephony is available in the vast majority of states as a relay service which is free to the user, (costs being paid by a levy on all telephone bills). Service users need a CapTel phone, costing approximately $400, but states either pay all or greatly subsidise this equipment cost. The US captioning service is provided by the parent

company of Teletec Ltd., called Ultratec Inc. – this CapTel service is usually offered under the Ultratec name, but is some areas the brand identity is that of other vendors such as Hamilton and Sprint. The American company offers CapTel and is preparing shortly to offer a web-carried captioning service, following the new UK WebCapTel model, (which was launched in April 2007).

The American caption service user places a call in the same way that a voice telephone user places a conventional phone call. As the user dials, the captioned telephone automatically connects to a captioning service. Call set-up is entirely invisible to the user; there is no interaction of any kind between the user and the communications assistant. This is different from other types of relay services; it is claimed that this "invisibility" empowers the user to make calls directly and to control the content and flow of the call. For example, if necessary, it is the captioned telephone user, not the communications assistant, who is the one who asks the speaker to repeat a word or spell a name.

Captioned telephone calls may also be initiated by non-captioned telephone users, though the manner of achieving this differs between the one- and two-line services. Individuals making incoming calls to a *one-line* captioned telephone user must first dial the toll-free captioning service and then enter the captioned telephone user's number in order for the user to receive captions of the conversation. Callers to a *two-line* captioned telephone user can simply dial the telephone number of the captioned telephone user, and the relay service for the captioned telephone is then connected automatically through the second telephone line. In either case, once the call is connected, the captioned telephone user will be able to hear the calling party and nearly simultaneously read captions of what the calling party is saying. Currently WebCaptel is like 1-line CapTel, but without the need for a special telephone. A hearing person wishing to call a WebCaptel user needs to dial Teletec's captioning centre first.

With whichever configuration, captioned telephone relay enables people with varying degrees of hearing loss to use (or more easily use) the telephone by displaying the text of the conversation on a special phone or on the screen of an Internet-connected device. The person with hearing loss voices their part of the conversation in the usual way and can read the text and/or listen to the incoming speech.

In June 2006, City University reported to Ofcom on its research into the "Feasibility of Additional Telephone Relay Services". The report contained a section entitled "Captioned telephony study" which was an examination of the current situation and certain aspects of market demand and technical potential.

Ofcom has since asked for a further study of captioned telephony, focusing on the following specific research questions –

Proprietary Technology

Would it be possible to design a captioned telephony service that does not rely on CapTel's proprietary technology? To what extent does the launch of WebCapTel offer a way around this obstacle?

Voice recognition

The original report described multiple speaker recognition as the holy grail of speech recognition. The availability of sophisticated voice recognition software would significantly reduce the costs of providing a captioned telephony service by obviating or diminishing the need for re-voicing. Given ICT developments which will provide enhanced random access memory and processing speed at ever-lower costs, how soon might we reasonably expect them to have an impact on captioned telephony services?

ScreenPhone

RNID's ScreenPhone offers some of the functionality of captioned telephony although reliant on a relay service operator rather than a re-voicing communications assistant. To what extent could this be considered an alternative to captioned telephony? Text communications will, by their nature, be slower than voice communications. Can we identify an acceptable delay beyond which alternative relay services cease to be a functional equivalent to voice telephony?

Alternative solutions

Are there other ways of providing the functionality of captioned telephony which have been developed elsewhere in the worlds or which technology is likely to enable within the medium term (3-5 years)?

PROPRIETARY TECHNOLOGY ISSUES

Captioned telephony in the UK

In the UK, Teletec's system for connecting the end user to the operator system differs from that of the RNID Typetalk/BT TextDirect system.

Teletec has a special network system, the details of which are kept commercially confidential with "proprietary technology". For the CapTel service this means that only Teletec handsets may be used with the Teletec system. The Teletec handsets do not use standard text communication protocols, so they cannot be connected to other relay services, or to other textphones.

The Teletec system allows voice and text to be sent together through the phone line, which the current BT TextDirect system cannot do. Teletec's proprietary system facilitates this simultaneous data traffic of voice and text on a single line, provided that the user is equipped with the CapTel phone. While voice and text data travel together on the phone line, they are not exactly synchronous because of the stages involved in transcribing the voice data into written text – these are: 1) the communications assistant hears the words being spoken by the hearing party in the phone call, 2) the assistant re-voices those words as promptly and as clearly as possible, 3) the assistant's computer recognises the assistant's words, makes a text version and transmits it to the hard of hearing or deaf person's screen on the Teletec CapTel phone or the Internet-connected device using WebCapTel. Teletec estimates that the time delay between the aural speech and the text display is between 3 and 5 seconds. Variation within this timing range or beyond may, commonsensically, depend on the performance of the individual communications assistant (in re-voicing speed for the particular accent and vocabulary as spoken by the hearing person) and on any variables in the processing of the equipment being used. We have found no public record of independent evaluation of Teletec or Ultratec captioning services' speed or text accuracy.

Using the CapTel service requires two types of payment – one for the purchase of the special Teletec phone hardware, the other for per-minute service use. The WebCapTel service requires no special equipment purchase but requires the per-minute service use charge since the web system connects to the Teletec network still using the company's proprietary system. This per-minute charge varies between 75p per minute (for a user who commits to £300 per month for total usage of 400 minutes) to £1.25 per minute (for a light user's captioning which exceeds a monthly £10 for 10 minutes). (http://www.webcaptel.co.uk/tariff.asp)

In the case of WebCapTel, the software-based proprietary system consists of the bundling and unpacking of voice-and-text data by control, signalling and media transport. There are existing open industry standards available for each of these elements of "mixed" data communication, but Teletec

uses and keeps confidential its particular combined system. Having been pioneers in developing telephone captioning services, the company seeks return on its investment.

Textphone standards and protocols

For textphones there are many different open standards. The original standard used by textphones is the Baudot code implemented asynchronously at either 45.5 or 50 baud, 1 start bit, 5 data bits, and 1.5 stop bits. Baudot is a common protocol in the US. In the UK, BT TextDirect / RNID Typetalk is a virtual V.18 network on the PSTN which offers interoperability between textphones using different protocols. The Ultratec / Teletec protocol is known as Enhanced TTY, which it calls "Turbo Code". The advantages of Turbo Code over Baudot protocol are that it gives a higher data rate, full ASCII compliance and full-duplex capability.

Textphones were developed before the rapid growth of the World Wide Web and Internet Protocol. Existing “legacy” PSTN textphones cannot function with simultaneous voice and text. On the one hand, the V.21 protocol uses continuous carrier (so that one switches between text and voice). On the other hand, the Baudot protocol lacks signalling for text/voice.

RNID has represented to us that, fundamentally, “VCO as such is a PSTN concept really”. In a purely IP environment these obstacles with textphones would not exist. However, the current UK and EU regulatory framework does not explicitly cover IP networks at all, so European national regulators may feel they have no authority over IP-based networks and services. It has been suggested to us, though, by organisations representing deaf people, that the existing laws could be interpreted as extending to IP access channels to traditional telecoms networks.

Present regulatory framework

The existing relay services legislative and regulatory framework is based on the PSTN and does not extend to IP networks. The Communications Act 2003 (and related secondary legislation) is based on the EU's 2002 framework for electronic communications, of which there are five founding Directives –

- Framework Directive (2002/21/EC) - on a common regulatory framework for electronic communications networks and services
- Access Directive (2002/19/EC) - on access to, and interconnection of, electronic communications networks and associated facilities
- Authorisation Directive (2002/20/EC) - on the authorisation of electronic communications networks and services
- Universal Service Directive (2002/22/EC) - on universal service and users' rights relating to electronic communications networks and services

- Privacy Directive (2002/58/EC) - on the processing of personal data and the protection of privacy in the electronic communications sector

The Universal Service Obligation Directive limits requirements as follows:

> *A fundamental requirement of universal service is to provide users on request with a connection to the public telephone network at a fixed location, at an affordable price. The requirement is limited to a single narrowband network connection, the provision of which may be restricted by Member States to the end-user's primary location/residence, and does not extend to the Integrated Services Digital Network (ISDN) which provides two or more connections capable of being used simultaneously.*

The connections must be made available to -

> *all end-users in their territory, irrespective of their geographical location, and, in the light of specific national conditions, at an affordable price.*

[Member states] may take special measures to –

> *ensure under the same conditions this access, in particular for the elderly, the disabled and for people with special social needs.*

Lacking explicit coverage to IP networks, national regulators such as Ofcom may not feel able to exert authority for IP-based networks and services. However, a regulator could form the view that where traditional telecoms networks interconnect with IP networks, the framework may enable regulation to extend to the IP access channel.

Specifically, the Access Directive establishes under Article 5 regulatory powers with regard to access and interconnection, which states –

> *National regulatory authorities shall, acting in pursuit of the objectives set out in Article 8 of Directive 2002/21/EC (Framework Directive), encourage and where appropriate ensure, in accordance with the provisions of this Directive, adequate access and interconnection, and interoperability of services, exercising their responsibility in a way that promotes efficiency, sustainable competition, and gives the maximum benefit to end-users.*

RNID has told us that it holds the view that "these powers are by themselves enough to enforce operators to make text telephony 'interconnect' and 'interoperate' between existing analogue textphones and VoIP environments - even if VoIP is as such out of scope for Ofcom at the moment."

At present there is no IP channel into RNID Typetalk and BT has now no current plans to provide this voluntarily. However, there are various IP-based real-time text solutions available, for example for Windows and mobile handsets so that IP technology could be used to offer captioned telephony through Typetalk.

IP-carried text relay service is available in the US from AT&T, Sprint, MCI and Hamilton Relay.

In the UK, it may be open to Ofcom to consider if it can or it should exercise powers from the Access Directive to prevail upon BT to provide an IP channel to TextDirect and to Typetalk in the interest of "adequate access", "interconnection" and "interoperability of services", particularly since Article 5 refers to giving "the maximum benefit to end-users".

IP interfacing with Typetalk

Several UK organisations have worked on IP interconnection with textphones. RNID is a very prominent researcher in this area and has also worked extensively on textphones' possible functional succession, recently launching TalkByText for Windows, Mobile Edition and Web Edition. The Windows version consists of software which emulates textphone hardware enabling a computer user to make and receive textphone calls through the PSTN or on VoIP. However, it does not have the Voice Carry Over option that users of textphones with handsets can take advantage of, being designed as a text-only textphone emulator (with additional features).

TalkByText Mobile Edition extends textphone functionality to mobile phones, being usable by a greater variety of mobiles than was the case with the RNID Mobile Textphone, and is capable of integration with IP-based real-time text. TalkByText Web Edition takes as much functionality as possible from the Windows version to be usable without the Windows version special software being installed, relying only on web browser capabilities in order to be used by somebody away from his or her own computer. "It is based on open industry standards, like SIP and integrates tightly with VOIP infrastructure". (http://www.ictrnid.org.uk/tbtwin.html) Currently TalkByText for Windows is only available as a business package but RNID New Technologies says that a home user version will be available in late 2007.

BT's views

From the BT side, we have been told that –

> *Unfortunately we have not done any research into providing support for simultaneous voice and text and we believe that the lack of suitable terminals for the end-users would make it very difficult to conduct research into this issue.*

This overall objection to the need for research might diminish as more people connect and interconnect between PSTN and VoIP. Even the initiative by RNID with TalkByText might grow the demand for "suitable terminals" while the general voice telephone market also inevitably leans towards VoIP.

The issue of interconnection, as distinct from discrete networks, further arises in BT continuing as follows –

> *For the support of end-users it would be necessary to design and build a service on top of an IP network. As you are probably aware BT provides TextDirect access to CPs so that their customers can use its facilities. For an IP solution to work, equivalent relationships would have to be established with IP connected end-users. To date no IP providers have expressed interest in providing support that would allow textphone users to use a native IP connection.*

Technically, BT also acknowledges that simultaneous voice and text on IP is simple, but it sees difficulties in service integration –

> *… Suppliers of IP based communication products should be urged to support voice and real-time text simultaneously which would in most cases only require a minor system change. If BT were approached by a provider of services that could support SVT [simultaneous voice and text] we would be willing to have commercial discussion about the facilitation of their customers' calls by TextDirect. If a dedicate service were to be provided as part of TextDirect this would be extremely costly, would not include additional facilities found on mainstream services, and would force textphone users into another service ghetto.*

Challenges to diversity in captioning services are regulatory and financial, not technical

Even with no regulatory changes, no further technical developments, no licensing of "proprietary technology", no new selection from existing open standards, a simple building-blocks simultaneous captioned telephony service could be created on a 2-line or triangular configuration –

- the hearing person is "conference"-connected on two lines, to both the deafened person and a communications assistant;
- the deafened person and the communications assistant are connected by a text device, such as a web-connected computer (with instant messaging) or a mobile phone (with texting);
- hearing person speaks – that voice is transcribed by the communications assistant (and may also be partially heard by the deafened person);
- deafened person speaks and is heard by the hearing person.

This simplified construction is feasible but there would be concerns about call signalling, control and connecting correctly. However, there are no real technical problems on either IP or PSTN (using open standards). The challenge would be funding, which connects with the legislative and regulatory definition of "relay service", a matter of secondary legislation.

Typetalk was originally funded by BT on a voluntary basis before becoming a mandatory obligation. There is neither a requirement for BT to do more than it currently performs in paying for RNID Typetalk, nor an incentive for them to take initiatives which will arouse further costs.

CapTel's success has been in focusing on the large captioning market and protecting their sales growth with proprietary technology which ties users to their particular equipment and service offering. CapTel has sold widely in the US where service use has been funded by the general levy on all telephone bills. CapTel's sales in the UK (undisclosed) have been very much more modest, since there is no comparable funding – nearly all current usage being paid for by Access to Work. WebCapTel is seeking to carry over the company's proprietary technology market-protective principle into IP and mobile domains.

RNID feels technically well able to expand into IP and PSTN-and-IP services but would need to persuade BT to fund something outside their legal obligation, or find funding from elsewhere.

The introduction of WebCapTel does not affect the possibility of designing a service unrelated to Captel's proprietary technology. The challenges to creating access to non-proprietary captioning are regulatory and financial rather than technical.

VOICE RECOGNITION

> *("Voice recognition" and "speech recognition" are widely used interchangeably as having identical meaning, as in this report. However, taken literally, "voice recognition" can have the meaning of identity recognition of a particular person, or gender recognition of whether a person is male or female. "Speech recognition" is the term taken precisely which needs no disambiguation to denote a conversion from meanings in spoken language to another form of data; [definition below]. We refer to "voice recognition" as the process of converting spoken words into text.)*

Voice recognition allows a computer to recognise words and follow basic vocal instructions by distinguishing phonemes (distinct sounds) and morphemes, the smallest units of linguistic meaning in a language.

A definition is that it is a computer's ability to "convert spoken speech into [text] data that it can then manipulate or execute" (Matthews, 2002). Similarly, Zue, Cole and Ward (1996) define it as "the process of converting an acoustic signal, captured by a microphone or telephone, into a set of words." As expressed in WordiQ (no date), it "allow[s] computers equipped with a source of sound input, such as a microphone, to interpret human speech, eg for transcription or as an alternative method of interacting with a computer."

The human voice produces an analogue signal. For use with computers, analogue audio must be converted into digital signals. This requires analogue-to-digital conversion. For a computer to decipher the signal, it must have a digital database, or vocabulary, of words or syllables, and a speedy means of comparing this data with signals. The speech patterns are stored on the hard drive and loaded into memory when the program is run. A comparator checks these stored patterns against the output of the analogue-to-digital converter.

In practice, the size of a voice recognition program's effective vocabulary is directly related to the random access memory (RAM) capacity of the computer in which it is installed. A voice recognition program runs many times faster if the entire vocabulary can be loaded into RAM, as compared with searching the hard drive for some of the matches. Processing speed is critical as well, because it affects how fast the computer can search the RAM for matches.

All voice recognition systems or programs make errors. Screaming children, barking dogs, and loud external conversations can produce false input. Much of this can be avoided only by using the system in a quiet room. There is also a problem with words that sound alike but are spelled differently and have different meanings, (such as "hear" and "here"), but this problem might eventually be largely overcome using stored contextual information (with cheaper RAM and faster processors).

The first software-only dictation product for PC's, Dragon Systems' Dragon Dictate for Windows 1.0, using discrete speech recognition technology, was released in 1994. Discrete speech is a slow, unnatural means of dictation, requiring a pause after each and every word. Two years later, IBM introduced the first continuous speech recognition software, its MedSpeak/Radiology. These systems were extremely costly and required very expensive PCs. Continuous speech technology allows its users to speak naturally and conversationally, relieving much of the tedium of discrete speech dictation.

Dragon Systems made an enormous stride in June 1997, when it released Naturally Speaking, the first general-purpose continuous speech software program. Much more affordable than earlier programs, it brought the realm of continuous speech recognition to a much wider range of users. Two months later, IBM released its competing continuous speech software, ViaVoice.

Computer process

Much is demanded of speech recognition programs. Accuracy is critical, and speed is essential to any effective program. Added to these challenges are the enormous variance that exists among individual human speech patterns, pitch, rate, and inflection. These variations are an extraordinary test of the flexibility of any program. Voice recognition follows these steps:

1. Spoken words enter a microphone.
2. Audio is processed by the computer's sound card.
3. The software discriminates between lower-frequency vowels and higher-frequency consonants and compares the results with phonemes, the smallest building blocks of speech. The software then compares results to groups of phonemes, and then to actual words, determining the most likely match.
4. Contextual information is simultaneously processed in order to more accurately predict words that are most likely to be used next, such as the correct choice out of a selection of homonyms such as “merry”, “marry”, and “Mary”.
5. Selected words are arranged in the most probable sentence combinations.
6. The sentence is transferred to a word processing application.

Voice recognition is much more challenging than voice synthesis (the latter producing clear sounds which can be understood as human spoken language). Because every person's voice is different, and words can be spoken in a range of different tones and emotions, the computational task of successfully recognising spoken words is very challenging, and has needed intensive specialist software research.

A variety of different approaches are used – dynamic algorithms, neural networks, and knowledge bases – with the most widely used underlying technology being the “Hidden Markov Model”, (a statistical modelling

method). These techniques all attempt to search for the most likely word sequence given the fact that the acoustic signal will also contain a lot of background noise. The task is made easier if the system can be trained to recognise one person's voice pattern rather than the varied patterns of many people, and it is also easier if isolated words are to be recognised rather than continuous voice. Similarly, the task is easier if the vocabulary is small, the grammar constrained and the context well-defined.

Grammar and context are particularly important elements in voice recognition, particularly in a highly complex language like English, and this has taken voice recognition system developers into areas like natural language analysis and comprehension.

The complexity of these problems has meant that most of the voice recognition systems developed to date are either small-vocabulary isolated-word recognition systems or large-vocabulary single-speaker recognition systems.

Parameter	Least robust systems	Most robust systems
Speaking mode	Isolated words	Continuous speech
Speaking style	Read speech	Spontaneous speech
User "flexibility"	Speaker dependent	Speaker independent
Vocabulary	Small (under 20 words)	Large (20,000 words or more)

According to these factors, the ideal speech recognition system would support fluent, continuous speech at a natural speaking rate; allow for a wide range of speakers (that is, its accuracy would not depend on training by one particular user); and give the speaker great flexibility in what he or she can say.

Commercial research directions for voice recognition

Although the technology for speaker-dependent large-vocabulary dictation systems now works quite well on a PC, they have not proved as popular in offices as was commercially predicted by software developers. Office workers include people for whom it is quicker and easier to edit a document using a conventional keyboard and mouse. Moreover, the high background noise levels found in some offices make recognition hard, and recognition rates can fall as low as 50 per cent compared with a normal quiet office level of up to 99 per cent.

The application of voice recognition has been more successful in telephony for "customer service"-type applications, not relay services. These are

applications that are not automatable using conventional push-button interactive voice response systems, such as directory assistance. Voice recognition technology is today widely used in automated phone-based information systems, such as travel booking and information, financial account information, and customer service call routing.

In such applications, accuracy of recognition is very high despite high noise levels, because such systems use constrained grammar recognition. This simply means that a highly optimised telephone application can trigger a prompt from the user to repeat the previous answer whenever the system's confidence in recognition of that input is low.

Voice recognition software is now increasingly used in mobile phones as a faster way to input SMS/text messages. Nuance Communications, one of the biggest producers of voice recognition products, claims that more than 50 million phones are now equipped with such software. For mobile phone users, although background noise levels can be very high, vocabulary size is much smaller and the grammar constrained, so once again recognition rates are high, (with short-duration voice training by the user, such as for 90 seconds).

In such applications, voice input is becoming popular because using a mobile phone can seem very time-consuming – because of multiple menus, options and sub-menu paths for each application. Just writing and sending a five-word SMS/text message can require as many as 50 keypad operations, while voice input allows the message to be input much more quickly than with a keypad.

Voice input of a mobile phone menu system also allows someone to use the phone without looking at the keypad – which is particularly appealing for car drivers.

In the US, another area where voice technology is growing in popularity is in voice-to-text dictation systems for use by professionals, such as doctors and lawyers. This is potentially a huge market with the healthcare area alone estimated to be worth more than $15bn (£8bn) annually – for the manual transcription of doctors' notes alone.

Messaging

The voice-to-text market is rapidly expanding in mobile and landline text messaging applications – for delivery of spoken messages in text form on mobile phones or by email. This market has become highly competitive between current service providers such as –

American companies – **VoiceSignal, SimulScribe, Jott, GotVoice, Copytalk, Phonewire, Dictomail**
New Zealand company – **Aangel**
UK companies – **MyJotter, SpinVox**
(All these companies have websites with URLs of their names plus ".com")

These providers vary in their techniques between human transcription and automatic speech recognition.

SpinVox is the leading UK service provider, expanding in the US and in French, Spanish and German language markets.

There is web chatroom and blog gossip about SpinVox using undeclared human transcribers, but the company assures us that their "Voice Message Conversion System" (VCMS) is automatic-operating software with humans only involved for vocabulary extension (of difficult words) and for quality control supervision.

The company's website makes no public declaration of its voice-to-screen techniques, which adds to web gossip, no doubt. Moreover, a SpinVox co-founder's patent (United States Patent 20060223502) specifies human transcription, but there may be any number of patents being used, untraceable at the time of this report and which are kept commercially confidential by the company.

SpinVox's VCMS does not provide real-time text/email messaging. A five second voicemail may take one minute to appear in its text form. The maximum message length possible on SpinVox's system is three minutes, which can take ten minutes for processing as text. The company claims a 97% text accuracy rate.

Challenges for phone captioning

In order to create accurate multi-speaker, wide-vocabulary, multi-location voice recognition for telephone captioning, the core challenges are the most demanding of general voice recognition development –

Homonyms and similar-sounding words

English and many other languages contain homonyms – words that have identical pronunciations but different definitions – and since these sound the same to a computer, they are often mistaken for the same word. Even words or phrases that do not sound exactly the same may be similar enough in pronunciation to confuse a speech recognition system. Even the most advanced designs may be destined never to capture all the richness of human languages (Kurzweil, 1996) but some such problems can be sidestepped by the use of accurate language models; (these are models which give a basic sense of context, grammar and "common sense" of meaning within language – language models usually define rules of grammar based on probability, rather than strict syntax requirements, allowing users to violate "correct" usage rules in their speech, so that the "common sense" likelihood that a particular word or pair of words will follow another word or pair can disambiguate between similar-sounding words, such as "there is" being much more common than "their is" in normal English speech).

Inter- and intra-speaker differences

Individual speakers of any language possess different accents, tones, speech impediments, vocal tract characteristics, and grammatical styles. Integrating all of these features into a single computer system is extremely difficult. Moreover, even the same speaker will often pronounce words differently, depending on his or her mood, inflection, and intention (e.g. sarcastic, serious, humorous). As WordiQ states, "Intonation and speech timbre can completely change the correct interpretation of a word or sentence, e.g. 'Go!', 'Go?' and 'Go.' can clearly be recognised by a human, but not so easily by a computer."

Importance of context and common sense

Having a basic measure of situational understanding is key to successful speech recognition. Often, humans only understand speech correctly because they can apply common sense developed *through human experience*, a skill which computers have yet to master. Kurzweil writes that "we understand speech in context. Spoken language is filled with ambiguities. Only our understanding of the situation, subject matter, and person (or entity) speaking – as well as our familiarity with the speaker – lets us infer what words are actually spoken." WordiQ concurs, noting that "words have different meanings in different sentences. 'Philip lies' could be interpreted either as Philip being a liar, or that Philip is lying on a bed."

Background noise / ambient sounds / audio interference

While this problem may seem the easiest to correct, it is perhaps the most permanent and pervasive. Whereas for example a WebCapTel communications assistant is in a specially quiet environment for his or her individually trained voice recognition computer, general speakers may not be observantly heedful to avoid extra sounds (or unable to prevent them). Background noise can be a significant problem. In addition, phone-based speech recognition may have to deal with poor transmission and static inherent to telephone lines, further decreasing its ability to capture the speaker's intent accurately.

Researchers are believed still to be years away from being able to produce a general purpose automatic voice recognition system that can recognise continuous voice from a wide variety of people and with a wide vocabulary as successfully as any human listener.

If every hearing person who wanted caption relay conversations with a hard-of-hearing or deaf person had previous software-training for voice recognition of their particular voices, then a service could be created with current technology – but this would be an organisational fantasy akin to science fiction.

The costs of random access memory and high-speed computer processors are not the main restraints on accurate multi-speaker, wide-vocabulary, multi-location voice recognition. It could be possible to organise a

telephone routing system whereby voice recognition calls could be diverted through a remotely-located computer with massive RAM and huge-speed processing. Such an arrangement would overcome the restrictive RAM and processing limits of domestic or ordinary office PCs. However, the essential linguistic and programming challenges remain, still postponing accurate multi-speaker, wide-vocabulary, multi-location voice recognition.

References

Kurzweil, R. (1996). When Will HAL Understand What We Are Saying? Computer Speech Recognition and Understanding. MIT Press. Accessed online 2 March 2007 – http://mitpress.mit.edu/e-books/Hal/chap7/seven1.html

Matthews, J. (2002). How Does Speech Recognition Work? 23 Oct. 2002. Generation5. Accessed online 5 March 2007 – http://www.generation5.org/content/2002/howsrworks.asp

WordiQ.com (no date). Definition of Speech Recognition. Accessed online 2 March 2007 – http://www.wordiq.com/definition/Speech_recognition

Zue, V., Cole R., & Ward W. (1996). Speech Recognition_1996. Oregon Graduate Institute of Science and Technology. Accessed online 4 March 2007 – http://cslu.cse.ogi.edu/HLTsurvey/ch1node4.html

SCREENPHONE

The RNID ScreenPhone, manufactured by Geemarc Telecom, was launched in October 2006. It is a new amplified telephone with the significant addition of a large screen for text relay display, when wanted. The development of the ScreenPhone was targeted to optimise ease of use. Its features are –

- Large screen 125mm x 80mm
- Adjustable text size on screen, up to 12mm
- One touch text and voice switching
- Flashing strobe light for incoming calls
- Adjustable ringer volume
- Adjustable receiving tone control
- Handset amplification increases to to 30dB @ 1KHz
- Text answerphone for receiving messages
- 50 number phonebook

It combines the functionality of an amplified phone with that of a textphone in a visually attractive form (according to taste). In its combined functionality, ScreenPhone can claim to be a new genre of telephone. The ScreenPhone is well-suited to Voice Carry Over (VCO) through RNID Typetalk (and to voice-only calls). It does not have a full alphabet keyboard for inputting whole words in TextDirect calls or non-VCO Typetalk relay calls, but an optional extra is a plug-in keyboard (which is relatively modestly priced, at £15.31, next to the much more substantial ScreenPhone cost of £199, [both prices excluding VAT]).

The ScreenPhone may be ideal for people who have age-related hearing loss, and may be existing or prospective users of RNID Typetalk. These people's choice (between ScreenPhone to be used with RNID Typetalk on the one hand and WebCapTel on the other) can be influenced by factors such as –

- familiarity with RNID Typetalk service
- affordability, (in view of the WebCapTel / CapTel captioning alternative costing between 75p and £1.25 per minute)
- home availability of Internet access needed for WebCapTel
- personal readiness in using Internet-connecting devices, as needed for WebCapTel

RNID estimates that there are probably approximately 450,000 people who cannot use voice telephones, even with amplification. The majority of these will have acquired their hearing impairment later in life. Many of these people like to use their own voices and may have difficulties or not want to type text for conversations, or may not want to use unfamiliar textphones.

Some of these people may be interested in using CapTel/WebCaptel.

However, nearly all current Teletec customers for CapTel and WebCapTel have Access to Work funding for the service costs. A high proportion of people with age-related hearing loss are retired people for whom Access to Work funding is unavailable.

We have sought views from some Hearing Concern members on the ScreenPhone. Their responses indicate that about £200 is both the expected general level and also the ceiling for a textphone, while a good amplified phone should be available at about £80. They confirm the view that using an assistive phone without needing to type is extremely welcome. They like the relay operator to be "invisible". However, a drawback of the ScreenPhone, intrinsic to RNID Typetalk, is having to push a button to listen or talk, and not being able to interrupt. They have drawn up a list of requirements for a relay service to be functionally equivalent to normal telephone usage:

- hear the other person – although they will not be able to understand everything.
- talk to the other person.
- have any intermediary not apparent to either speaker or listener.
- have the text presented, as far as possible, synchronous with the voice.
- be able to interrupt if necessary.
- have callers dial a number and start the call with automatic inclusion of the relay but with minimal knowledge that the call is via a relay.
- be able to call emergency numbers and receive responses like hearing callers.

The ScreenPhone fulfils most but not all of these requirements. The issue of synchronicity between voice and text is not relevant for a Typetalk-related phone, though it is a crucial issue with CapTel/WebCapTel, where voice and text are used together. Hearing Concern members wanted the delay between voice and text as short as possible but thought that people had to be realistic:

> *If there is an intermediary who is speaking the hearing user's voice into the speech recognition system for delivery to the hard of hearing user as text, there is likely to be a delay at least as long as 400 msec even before the voice recognition system gets going. I think we need to be realistic about this and accept that ideally the delay needs to be kept as short as possible, preferably less than a second but until some meaningful tests have been done I'm not sure that we have enough data to give a better recommendation.*

Another Hearing Concern member thought the 3-5 second delay estimated for WebCapTel quite acceptable.

ALTERNATIVE SOLUTIONS

Captioned telephony is a system of providing synchronized text displays of spoken language in a telephone conversation, so that a person can read transcripted words to replace or to augment hearing and understanding the spoken language.

Functionally, captioned telephony provides written words for speech.

It should do so –

- quickly, as close to simultaneous as possible
- accurately

Desirable further attributes are that –

- method should be as unintrusive as possible for a near-natural conversation
- method should be easy to operate

To provide the functionality of voice recognition captioned telephony, a service does need to generate text quickly. One of the problems with traditional text relay is that it is relatively slow. In general, natural, brisk speech, people talk at 150 to 200 words a minute, but a top-skilled typist may be able to maintain half that speed for a protracted time. So it generally takes considerably longer for a communications assistant or relay operator to type a spoken message than it took for the hearing person to say it. Voice recognition offers the possibility of having the typed text nearly keep up with the spoken message.

However, as confirmed in this research, accurate voice recognition software must be trained on the voice of each speaker. It can't "listen" to the hearing person and accurately convert that speech to text without being trained to the individual voice. Instead, the software is trained on the voice of the communications assistant, who repeats what the hearing person says into the voice recognition system, which converts it to text.

Functional similarity

The nearest functional rival to captioned telephony is two-line VCO. Whereas in one-line VCO, the PSTN systems' protocols only allow either voice or text on the single telephone line at any one time, two-line VCO allows the lines to carry both voice and text at the same time. One-line VCO has the additional relative disadvantage that the VCO user cannot hear the other person speak, as he or she might wish to if residual hearing is sufficient.

Two-line VCO calls also give the service user more control over their calls than by single-line "traditional" text relay because the communication assistant does not identify the relay and is present only to type the voice of the called party: it is an "invisible" system without the need for signals like "GA" and "SK" to cue each person and to end the call. To take

advantage of these features of two-line PSTN VCO, the user must have two separate phone lines with different numbers and 3-way or conference calling capability on one of those lines.

A short online Quicktime movie illustrating the merits of two-line VCO is available from the US telephone company Sprint as service provider for Relay North Carolina at -- http://www.relaync.com/movies/2vcoQT.htm

Two-line VCO is functionally closest to captioned telephony of speech-text relay options. Two-line VCO can perform very well by virtue of highly accurate human understanding of spoken words and fast typing. Captioned telephony with voice recognition may have accuracy mistakes but is likely to perform faster for longer, although the re-voicing performance of the individual communications assistant must bear upon the pace and accuracy of the conversation text. There are no independent, published, performance assessments of captioned telephony that we have been able to trace.

Internet variant

While two-line PSTN VCO requires the user to have two separate phone lines with different numbers and 3-way or conference calling on one of the lines, there is also an Internet variant operating in the US (since launch in July 2004 of AIM Relay Services) –

> User needs an Internet connection and AIM (AOL Instant Messenger) and a single landline telephone line with 3-way or conference calling.
>
> *Setup instructions to make a 2-line VCO call:*
>
> 1. sign into AIM and send an IM to a relay operator.
> 2. tell the relay operator "I want to make a 2-line VCO call" and give the operator your landline number
> 3. the operator calls your number – ask the operator "can you hear me?"
> 4. the operator confirms by typing it in the IM box so you can read it
> 5. once this is confirmed you tell the operator to hold while you dial the number of the person you want to call.
> 6. when that person answers, you connect all three (yourself, operator and person you are calling) into a conference call
> 7. you talk into your phone to the person you are calling – when the called person replies the operator types that to you in the IM box.

The AIM service is also available through wireless devices and mobile phones with WAP-enabled Internet access.

Visual options

Other telephone relay innovations have been developing non-textual visual systems. For example Synface, (which involved British, Dutch and Swedish researchers in a project developed from Swedish Teleface), has experimented with a screen representation of a lip-readable speaking face. A Swedish company, SynFace, is now offering EyePhone software with which a user can see a synthesized talking face for real-time lip-reading from a computer screen.

Also in Sweden, the Omnitor company has devised Allan eC, standing for "All languages electronic Conversation", a multifunction-terminal system for sign language, text and voice for “all in one” communications diversity. Allan eC follows international standards for the concept of Total Conversation with capabilities for simultaneous video, text and voice through the one multifunctional terminal. In each conversation, the communicating parties can use text and/or voice and/or video. The purpose of using any of the three media is either to converse directly, or to provide some additional communication which supports the conversation. Allan eC is a single terminal which can be used for whichever form of language communication is chosen by the people on the call – the one piece of hardware enables dialogues in vision for sign language, in text and/or by voice. Omnitor acknowledges that re-voicing and voice recognition may be the best way to produce text versions of spoken language, but expresses strong support for creating relay systems only with open international standards (for interoperability, access to emergency links and for general optimal service development).

Omnitor also observes that the Synface approach to synthesized faces for lip-reading could also possibly lead to automatic phonetic text-language generation, which would be readable with some training.

ACKNOWLEDGEMENTS

We are very grateful for the generous amount of time and assistance given to us for this project by RNID, Teletec International, Hearing Concern, BT and Omnitor. Of course, the responsibility for any errors or omissions, and for interpretation of the situation is entirely ours.