

Speech Recognition for Smart Homes

Ian McLoughlin and Hamid Reza Sharifzadeh
*Nanyang Technological University
Singapore*

1. Introduction

When Christopher Sholes created the QWERTY keyboard layout in the 1860s (often assumed to be for slowing down fast typists), few would have imagined that his invention would become the dominant input device of the 20th century. In the early years of the 21st century (the so called 'speed and information' century), its use remains dominant, despite many, arguably better, input devices having been invented. Surely it is time to consider alternatives, in particular the most natural method of human communications - spoken language.

Spoken language is not only natural, but in many cases is faster than typed, or mouse-driven input, and is accessible at times and in locations where keyboard, mouse and monitor (KMM) may not be convenient to use. In particular, in a world with growing penetration of embedded computers, the so-called 'smart home' may well see the first mass-market deployment of vocal interaction (VI) systems.

What is necessary in order to make VI a reality within the smart home? In fact much of the underlying technology already exists - many home appliances, electrical devices, infotainment systems, sensors and so on are sufficiently intelligent to be networked. Wireless home networks are fast, and very common. Speech synthesis technology can generate natural sounding speech. Microphone and loudspeaker technology is well-established. Modern computers are highly capable, relatively inexpensive, and - as embedded systems - have already penetrated almost all parts of a modern home. However the bottleneck in the realisation of smart home systems appears to have been the automatic speech recognition (ASR) and natural language understanding aspects.

In this chapter, we establish the case for automatic speech recognition (ASR) as part of VI within the home. We then overview appropriate ASR technology to present an analysis of the environment and operational conditions within the home related to ASR, in particular the argument of restricting vocabulary size to improve recognition accuracy. Finally, the discussion concludes with details on modifications to the widely used Sphinx ASR system for smart home deployment on embedded computers. We will demonstrate that such deployments are sensible, possible, and in fact will be coming to homes soon.

2. Smart Homes

The ongoing incorporation of modern digital technology into day to day living, is likely to see smart homes joining the next wave of computational technology penetration (McLoughlin & Sharifzadeh, 2007). This is an inevitable step in the increasing convenience

and user satisfaction in a world where users expect to be surrounded and served by many kinds of computers and digital consumer electronics products.

In parallel to this, advancements in networking have led to computer networks becoming common in everyday life (Tanenbaum, 1996) – driven primarily by the Internet. This has spawned new services, and new concepts of cost-effective and convenient connectivity, in particular wireless local-area networks. Such connectivity has in turn promoted the adoption of digital infotainment.

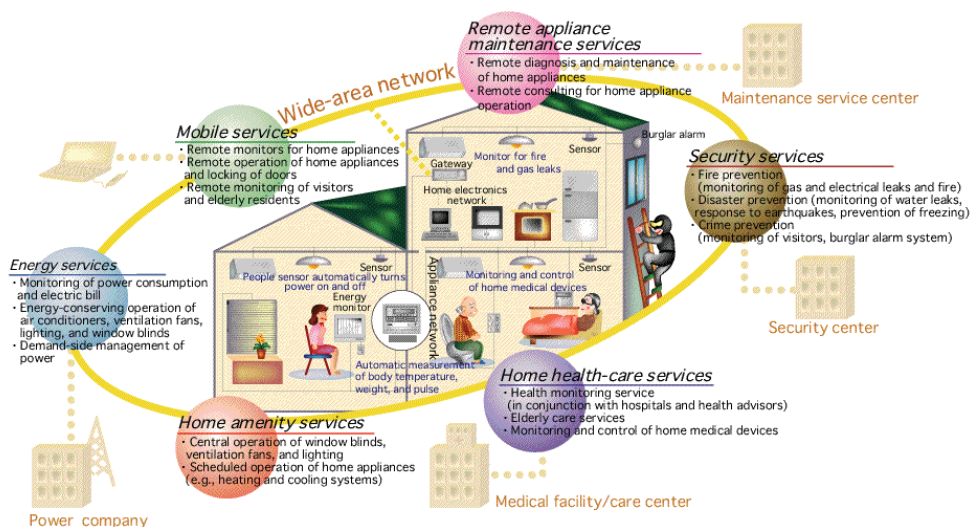


Fig. 1. An illustration of the range and scope of potential smart home services, reproduced by permission of ECHONET Consortium, Japan (ECHONET, 2008).

Recently trends reveal that consumers are more often buying bundles of services in the area of utilities and entertainment, while technical studies in the field of connected appliances (Lahrman, 1998; Kango et al., 2002b) and home networking (Roy, 1999) are showing increasing promise, and increasing convergence in those areas. Figure 1 illustrates many of the services that can be provided for various activities within a house (ECHONET, 2008). An appliance can be defined as smart when it is 'an appliance whose data is available to all concerned at all times throughout its life cycle' (Kango et al., 2002). As a matter of fact, smart appliances often use emerging technologies and communications methods (Wang et al., 2000) to enable various services for both consumer and producer.

Here we define smart homes as those having characteristics such as central control of home appliances, networking ability, interaction with users through intelligent interfaces and so on. When considering natural interaction with users, one of the most user-friendly methods would be vocal interaction (VI). Most importantly, VI matches well the physical environment of the smart home. A VI system that can be accessed in the garage, bathroom, bedroom and kitchen would require at least a distributed set of microphones and loudspeakers, along with a centralised processing unit. A similar KMM solution will by contrast require keyboard, mouse and monitor in each room, or require the user to walk to a centralised location to perform input and control. The former solution is impractical for cost

and environmental reasons (imagine using KMM whilst in the shower), the latter solution is not user-friendly.

Practical VI presupposes a viable two way communications channel between user and machine that frees the user from a position in front of KMM. It does not totally replace a monitor – viewing holiday photographs is still more enjoyable with a monitor than through a loudspeaker – and in some instances a keyboard or mouse will still be necessary: such as entering or navigating complex technical documents. However a user-friendly VI system can augment the other access methods, and be more ubiquitous in accessibility, answering queries and allowing control when in the shower, whilst walking up stairs, in the dark and even during the messy process of stuffing a turkey.

The following sections focus on ASR issues as an enabling technology for VI in smart home computing, beginning with an overview of ASR evolution and state-of-the art.

3. ASR development

Half a century of ASR research has seen progressive improvements, from a simple machine responding to small set of sounds to advanced systems able to respond to fluently spoken natural language. To provide a technological perspective, some major highlights in the research and development of ASR systems are outlined:

The earliest attempts in ASR research, in the 1950s, exploited fundamental ideas of acoustic-phonetics, to try to devise systems for recognizing phonemes (Fry & Denes, 1959) and recognition of isolated digits from a single speaker (Davis et al., 1952). These attempts continued in the 1960s by the entry of several Japanese laboratories such as Radio Research Lab, NEC and Kyoto University to the arena. In the late 1960s, Martin and his colleagues at RCA Laboratories developed a set of elementary time-normalisation methods, based on the ability to reliably detect the presence of speech (Martin et al., 1964). Ultimately he founded one of the first companies which built, marketed and sold speech recognition products.

During the 1970s, speech recognition research achieved a number of significant milestones, firstly in the area of isolated word or discrete utterance recognition based on fundamental studies by in Russia (Velichko & Zagoruyko, 1970), Japan (Sakoe & Chiba, 1978), and in the United States (Itakura, 1975). Another milestone was the genesis of a longstanding group effort toward large vocabulary speech recognition at IBM. Finally, researchers in AT&T Bell Laboratories initiated a series of experiments aimed at making speech recognition systems that were truly speaker independent (Rabiner et al., 1979). To achieve this goal, sophisticated clustering algorithms were employed to determine the number of distinct patterns required to represent all variations of different words across a wide population of users. Over several years, this latter approach was progressed to the point at which the techniques for handling speaker independent patterns are now well understood and widely used.

Actually isolated word recognition was a key research focus in the 1970s, leading to continuous speech recognition research in the 1980s. During this decade, a shift in technology was observed from template-based approaches to statistical modelling, including the hidden Markov model (HMM) approach (Rabiner et al., 1989). Another new technology, reintroduced in the late 1980s, was the application of neural networks to speech recognition. Several system implementations based on neural networks were proposed (Weibel et al., 1989).

The 1980s was characterised by a major impetus to large vocabulary, continuous speech recognition systems led by the US Defense Advanced Research Projects Agency (DARPA) community, which sponsored a research programme to achieve high word accuracy for a thousand word continuous speech recognition database management task. Major research contributions included Carnegie-Mellon University (CMU), inventors of the well known Sphinx system (Lee et al., 1990), BBN with the BYBLOS system (Chow et al., 1987), Lincoln Labs (Paul, 1989), MIT (Zue et al., 1989), and AT&T Bell Labs (Lee et al., 1990).

The support of DARPA has continued since then, promoting speech recognition technology for a wide range of tasks. DARPA targets, and performance evaluations, have mostly been based on the measurement of word (or sentence) error rates as the system figure of merit. Such evaluations are conducted systematically over carefully designed tasks with progressive degrees of difficulty, ranging from the recognition of continuous speech spoken with stylized grammatical structure (as routinely used in military tasks, e.g., the Naval Resource Management task) to transcriptions of live (off-the-air) news broadcasts (e.g. NAB, involving a fairly large vocabulary over 20K words) and conversational speech.

In recent years, major attempts were focused on developing machines able communicate naturally with humans. Having dialogue management features in which speech applications are able to reach some desired state of understanding by making queries and confirmations (like human-to-human speech communications), are the main characteristics of these recent steps. Among such systems, Pegasus and Jupiter developed at MIT, have been particularly noteworthy demonstrators (Glass & Weinstein, 2001), and the How May I Help You (HMIHY) system at AT&T has been an equally noteworthy service first introduced as part of AT&T Customer Care for their Consumer Communications Services in 2000 (Gorin, 1996). Finally, we can say after almost five decades of research and many valuable achievements along the way (Minker & Bennacef, 2004), the challenge of designing a machine that truly understands speech as well as an intelligent human, still remains. However, the accuracy of contemporary systems for specific tasks has gradually increased to the point where successful real-world deployment is perfectly feasible.

4. ASR in smart homes

Speech recognition applications can be classified into three broad groups (Rabiner, 1994) of isolated word recognition systems (each word is spoken with pauses before and afterwards, such as in bank or airport telephony services), small-vocabulary command-and-control applications, and large vocabulary continuous speech systems.

From an ASR point of view, a smart home system would aim to be a mixture of the second and third classes: predefined commands and menu navigation can be performed through a grammar-constrained command-and-control vocabulary, while email dictation and similar applications would involve large vocabulary continuous speech recognition. By and large, we can classify an ASR system in a smart home through its vocal interaction in two main categories: first are specific control applications which form the essence of smart homes, and second are general vocal applications which any ASR systems can carry out. We can summarize these categories and their characteristics as the following:

4.1 Smart home control

Probably the main feature of a smart home is the ability to vocally command the functions of the home and its appliances. We refer to this as the command-and-control function.

Most command-and-control applications have a small vocabulary size (0 to 50 words), reflecting the operations required to control the equipment. For example, commands for controlling the lights might include 'on', 'off', the location, and perhaps a few more words, depending on what additional operations are available. In addition the device being controlled should be identified (for example 'please turn on the bathroom light' is made up of a framing word 'please', an operation 'turn on', a location 'bathroom' and a device 'light'). Usually there is a direct mapping between the word or phrase and its semantics, i.e. the action to be carried out or the meaning to be associated with the words. However, more complex commands can be managed through a set of alternatives, where the vocabulary is restricted and known, such as week days or times of the day. As the number of alternative wordings increases, the task of listing all possible combinations and associating them with a given set of actions become unmanageable and so a grammar syntax is required that specifies, in a more abstract way, the words and phrases along with their permissible combinations.

4.2 Vocal interaction for information access

It is the view of the authors that for now, and some time to come, the Internet is likely to constitute the most common route for information access, alongside the stored files and archive of particular users. However it is accessing the vast and diverse World Wide Web (WWW) that is likely to post the most technically challenging tasks for VI. It is feasible to assume that the predominant graphical/textual nature of the current WWW is a natural consequence of the graphical/textual bias of HTML, which is most of all due to the way in which users are conditioned to interact with computers through KMM. If users commonly interacted with the WWW in a vocal fashion then it is quite possible that voice-enabled 'pages' would appear. To date this has not been the case.

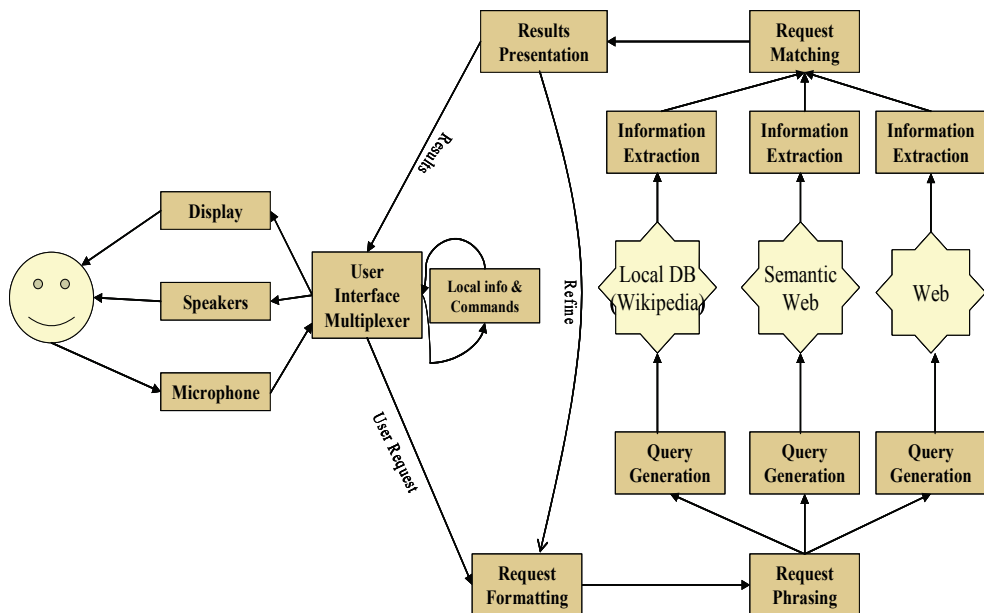


Fig. 2. Overall structure of the WWW vocal query access system.

The semantic web currently being promoted and researched by Tim Berners-Lee and others (see Wikipedia 2008), goes a long way towards providing a solution: it divorces the graphical/textual nature of web pages from their information content. In the semantic web, pages are based around information. This information can then be marked up and displayed *graphically* if required. When designing smart home services benefiting from vocal interactions of the semantic web, the same information could be marked up and presented vocally, where the nature of the information warrants a vocal response (or the user requires a vocal response).

There are three alternative methods of VI relating to the WWW resource:

- The few semantic web pages (with information extracted and then, either as specified in the page, or using local preferences, converted to speech), and then presented vocally.
- HTML web pages, with information extracted, refined then presented vocally.
- Vocally-marked up web pages, presented vocally.

Figure 2 shows the overall structure proposed by the authors for vocal access to the WWW. On the left is the core vocal response system handling information transfer to and from the user. A user interface and multiplexer allow different forms of information to be combined together. Local information and commands relate to system operation: asking the computer to repeat itself, take and replay messages, give the time, update status, increase volume and so on. For the current discussion, it is the ASR aspects of the VI system which are most interesting:

User requests are formatted into queries, which are then phrased as required and issued simultaneously to the web, the semantic web and a local Wikipedia database. The semantic web is preferred, followed by Wikipedia and then the WWW.

WWW responses can then be refined by the local Wikipedia database. For example too many unrelated hits in Wikipedia indicate that query adjustments may be required. Refinement may also involve asking the user to choose between several options, or may simply require rephrasing the question presented to the information sources. Since the database is local, search time is almost instantaneous, allowing a very rapid request for refinement of queries to be put to the user if required before the WWW search may have completed.

Finally, results are obtained as either Wikipedia information, web pages or semantic information. These are analysed, formatted, and presented to the user. Depending on the context, information type and amount, the answer is either given vocally, graphically or textually. A query cache and learning system (not shown) can be used to improve query processing and matching based on the results of previous queries.

4.3 Dictation

Dictation involves the automatic translation of speech into written form, and is differentiated from other speech recognition functions mostly because user input does not need to be interpreted (although doing so may well aid recognition accuracy), and usually there is little or no dialogue between user and machine.

Dictation systems imply large vocabularies and, in some cases, an application will include an additional specialist vocabulary for the application in question (McTear, 2004). Domain-specific systems can lead to increased accuracy.

4.4 Open-ended commands

Imagine a computer-based butler. This would need to be able to interpret, understand, and carry out complex commands. For example 'please book me a night flight to Tokyo next Thursday and reserve a suitable suite at the Grand Hyatt Hotel', or perhaps, 'please check my address book and make a dental appointment at an appropriate time over the next few days'. There is not only a significant vocabulary size implied in such commands, but also the understanding of the meaning of the command, and an appreciation of related factors: for example the words 'suitable' and 'appropriate' demand value judgments.

Quite clearly these commands impinge on the area of natural language processing (NLP) research, but do represent the ultimate destination of research into smart home systems – the home which is able to conveniently cater to our needs. It is likely to be several years before such open-ended commands can be handled successfully by automated systems.

4.5 General ASR applications

Since the advent of computers, the need to collect and neatly present documents has required textual data entry. The potential of ASR systems is that much of this process can be performed through VI. Many data entry applications involve predefined items such as name and address, for example form completion, package sorting, equipment maintenance, and traffic accident reports.

Data entry applications usually have limited vocabulary size including numbers, name and address details, and several additional control words. However, there may also be a requirement for a significant number of application-specific words, depending on the application type (McTear, 2004).

Similarly, computer games will likely form potential future applications which can only be guessed at currently. Educational use also had great potential – much teaching is conveyed vocally in schools and universities worldwide, and an individual education for children, adapting to their pace and needs, may well become a reality with advanced VI systems.

Finally, most societies contain people battling with loneliness, nobody to talk to, nobody who understands them. The thought of a machine companion may seem far-fetched today, but in the opinion of the authors it is only a matter of time before a sufficiently powerful and responsive computer system could become a best friend to some in society. As with the entire smart home project, what is needed is a sufficiently natural and accurate VI system coupled with a computer system that could pass the Turing test, at least when conducted by their potential clients.

5. Vocabulary size and performance

The ability of a system to recognise captured speech, to cater for intra- and inter-speaker variability, and the processing time allowable for recognising utterances are three main usability issues related to VI systems. Other issues include training requirements, robustness, linguistic flexibility and dialogue interaction.

Many factors in ASR for VI can be controlled. For example the variability of speech is mostly confined to a limited set of uses: linguistic flexibility can, and should, be constrained through appropriate grammar design (which is focus of section 6) and so on. The ability to accurately recognize captured speech that has been constrained in the ways discussed above will then depend primarily upon vocabulary size and speech-to-noise ratio. Thus we can

improve recognition firstly by restricting vocabulary size, and secondly by improving signal-to-noise ratio. The former task, constraint of vocabulary size, is the role of constructed grammar in VI systems.

It is well known that vocabulary restrictions can lead to recognition improvements whether these are domain based (Chevalier et al., 1995) or simply involve search-size restriction (Kamm et al., 1994). Similarly the quality of captured speech obviously affects recognition accuracy (Sun et al., 2004). Real-time response is also a desirable characteristic in many cases.

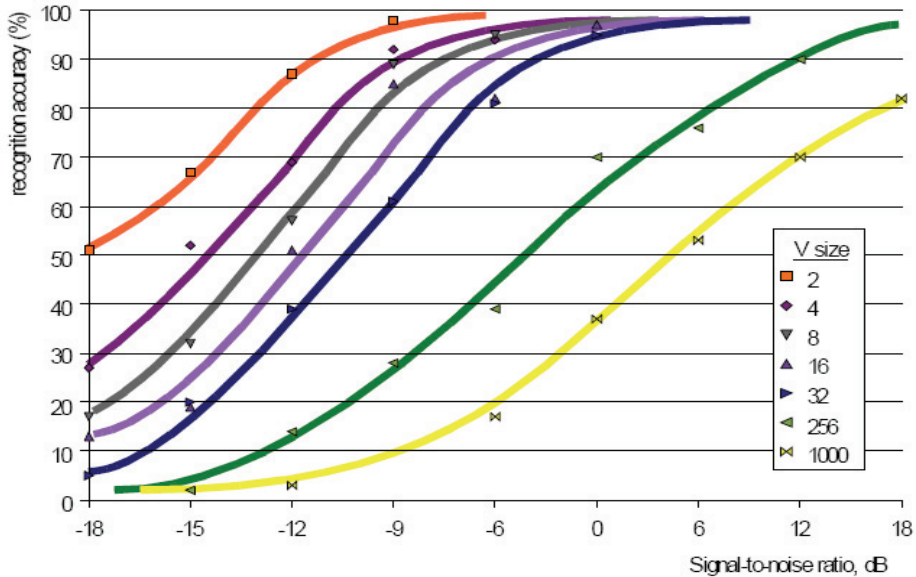


Fig. 3. Effect of vocabulary size and SNR on word recognition by humans, after data obtained in (Miller et al, 1951).

Actually the three aspects of performance: recognition speed, memory resource requirements, and recognition accuracy, are in mutual conflict, since it is relatively easy to improve recognition speed and reduce memory requirements at the expense of reduction in accuracy (Ravishankar, 1996). The task for designing a vocal response system is thus to restrict vocabulary size as much as practicable at each point in a conversation. However, in order to determine how much the vocabulary should be restricted, it is useful to relate vocabulary size to recognition accuracy at a given noise level.

Automatic speech recognition systems often use domain-specific and application-specific customisations to improve performance, but vocabulary size is important in any generic ASR system regardless of techniques used for their implementation.

Some systems have been designed from the ground-up to allow for examination of the effects of vocabulary restrictions, such as the Bellcore system (Kamm et al., 1994) which provided comparative performance figures against vocabulary size: it sported a very large but variable vocabulary of up to 1.5 million individual names. Recognition accuracy decreased linearly with logarithmic increase in directory size (Kamm et al., 1994).

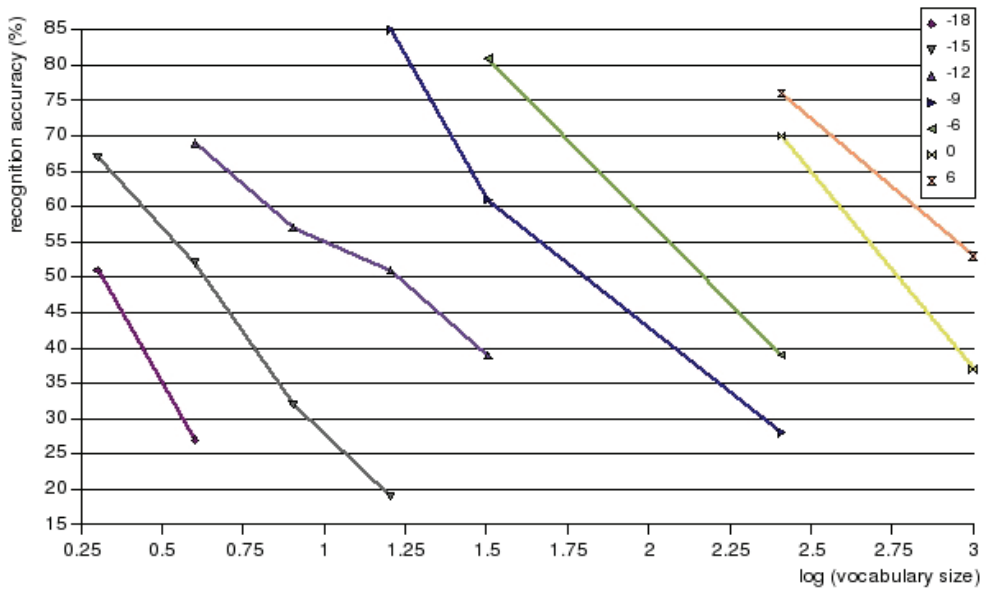


Fig. 4. Plot of speech recognition accuracy results showing the linear decrease in recognition accuracy with logarithmic increase in vocabulary size in the presence of various levels of SNR (McLoughlin, 2009).

To obtain a metric capable of identifying voice recognition performance, we can conjecture that, in the absence of noise and distortion, recognition by human beings describes an approximate upper limit on machine recognition capabilities: the human brain and hearing system is undoubtedly designed to match closely with the human speech creation apparatus. In addition, healthy humans grow up from infancy with an in-built feedback loop to match the two.

While digital signal processing systems may well perform better at handling additive noise and distortion than the human brain, to date computers have not demonstrated better recognition accuracy in the real world than humans. As an upper limit it is thus instructive to consider results such as those from Miller et al. (Miller et al, 1951) in which human recognition accuracy was measured against word vocabulary size in various noise levels.

The graph of figure 3 plots several of Miller’s tabulated results (Kryter, 1995), to show percentage recognition accuracy against an SNR range of between -18 and +18dB_{SNR} with results fit to a sigmoid curve tapering off at approximately 100% accuracy and 0% accuracy at either extreme of SNR. Note that the centre region of each line is straight so that, irrespective of the vocabulary, a logarithmic relationship exists between SNR and recognition accuracy. Excluding the sigmoid endpoints and plotting recognition accuracy against the logarithm of vocabulary size, as in figure 4, clarifies this relationship (McLoughlin, 2009).

Considering that the published evidence discussed above for both human and computer recognition of speech shows a similar relationship, we state that in the presence of moderate levels of SNR, recognition accuracy (A) reduces in line with logarithmic increase in

vocabulary size (V), related by some system dependent scaling factor which will represent by γ :

$$A^{-1} = \gamma \log(V) \quad (1)$$

6. Grammar structures for smart homes dialogues

As seen in section 4, VI for smart homes involves different levels of complexity ranging from an isolated word recognizer to an unconstrained ASR dictation system. For isolated word recognition, a speech recognizer attempts to detect commands by focusing on keywords in segmented natural sentences from the talker. A more complicated system, often referred to as a system driven dialogue (Nakano et al., 2006), can allow a user to complete information fields. Large vocabulary continuous speech recognition (LVCSR), by contrast, requires language models or grammars to select the most likely word sequence from the relatively large number of alternative word hypotheses produced during the search process at each stage in a conversation. Simple recognition tasks can use rule-based regular or context free grammars, where the system only recognizes a limited vocabulary.

In a speech-based smart home system, a VI session begins with a configurable spoken identifier phrase to 'attract the attention' of the ASR computer. From this point on, the adoption of grammar and syntactic structures, although these must be learnt by the user and thus reduce user friendliness, are crucial in maintaining the recognition accuracy of such systems.

Following attention, a tree of vocabulary options is possible. The aim at each stage of is to maximise recognition accuracy for given SNR, by reducing the vocabulary search space. An example sentence for one speaker might be:

ACTION - argument - (optional MODIFIER- (optional argument)) - PAUSE/REPEAT

In general, each of the action, argument and modifier sub phrases will have differing vocabulary characteristics, but it is a common feature of such systems that vocabulary size be minimised through syntactic structure. Let us proposed that each sub-phrase is classified by the following arguments:

- *Vocabulary consisting of V elements*
- *Accuracy requirement, R*
- *Length constraint, L*
- *Interruptibility type, T*

The characteristics of each sub-phrase are known in advance, having been defined by the grammar syntax, with recall prompted by traversal to the current branch. The length constraint is used to help detect end-of-phrase overrun conditions, and perhaps a missed MODIFIER phrase.

One or more phrases by each party in the communication comprise a conversation (or dialogue). An overall conversation, and indeed a phrase made of sub phrases has its own accuracy requirement. However this level of end-to-end link assurance is unwieldy for anything but confirming that an important action should or should not occur as the result of a conversation.

Sub-phrase vocabulary may occasionally need to be unconstrained (such as when performing a web search or dictating an email), but at other times could be limited (in this mode, we can use keywords or embedded phrases in natural sentences to command the

system). Potentially these two classes of recognition task could be performed with two different types of ASR software employing isolated word and continuous speech recognition respectively (Nakano et al., 2006). However both are catered for in Sphinx2 using two density acoustic models, namely semi-continuous and continuous. Some other flexible speech recognition systems have been introduced.

One system by Furui (Furui, 2001) uses a method of automatically summarizing speech, sentence by sentence. This is quite domain-specific (closed-domain) with limited vocabulary size (designed for news transcription), and may not be directly applicable to a continuously variable vocabulary size smart home system. However it does perform very well, and provides an indication of the customisations available in such systems.

Vocabulary size, V , impacts recognition accuracy, and needs to be related to accuracy requirement, R . Since 100% accuracy is unlikely, smart home systems need to be able to cope with inaccuracies through careful design of the interaction and confirmation processes. In particular, a speech recognizer that provides a confidence level, C can tie in with sub-phrase arguments to determine requests-for-clarification (RFC) which are themselves serviced through examination of the interruptibility type, T .

So given a recognition confidence level C , an RFC will be triggered if:

$$\frac{C\gamma}{\log(V)} < R \quad (2)$$

Where γ is system and scale dependent, determined through system training. Interruptibility type includes two super-classes of 'immediate' and 'end-of-phrase'. Immediate interrupts may be verbal or non-verbal (a light, a tone, a gesture such as a raised hand, or a perplexed look on a listeners' face based on the designed interface). An immediate interrupt would be useful either when the utterance is expected to be so long that it is inconvenient to wait to the end, or when the meaning requires clarification up-front. An example of an immediate interrupt would be during an email dictation, where the meaning of an uncertain word needs to be checked as soon as the uncertainty is discovered – reviewing a long sentence that has just been spoken in order to correct a single mistaken word is both time consuming and clumsy in computer dialogue terms.

An end-of-phrase interrupt is located at a natural reply juncture, and could be entirely natural to the speaker as in "did you ask me to turn on the light?"

7. Embedded speech recognition

Nowadays, embedded speech technology as an active research area attracts not only researchers from academia but also industrial groups interested to invest in this promising new market. Thus, more and more companies have launched embedded speech systems. These provide alternative control interfaces for consumer appliances to replace knobs, switches, buttons and so on. In specific niche applications, with limited vocabulary size, the success of such niche products may well advance the public acceptance of speech technology. Current examples include voice dialling for GSM telephones, and media players.

As consumer devices become increasingly complex, naturally the range of features increases, and thus it has become more and more difficult for users to produce the appropriate sequences of key presses to set a control. A typical example is the inability of

most people to use a remote control to set the timer on their video recorder to record forthcoming broadcasts. In addition, as devices decrease in size, and average users increase in age, manual manipulation has similarly become more difficult. From a system architecture point of view, embedded speech recognition is now becoming considered a simple approach to user interfacing. Adoption in the embedded sphere contrasts with the more sluggish adoption of larger distributed system approaches (Tan & Varga, 2008).

However there is a price to be paid for such architectural simplicity: complex speech recognition algorithms must run on under-resourced consumer devices. In fact, this forces the development of special techniques to cope with limited resources in terms of computing speed and memory on such system.

Resource scarcity limits the available applications: on the other hand it forces algorithm designers to optimise techniques in order to guarantee sufficient recognition performance even in adverse conditions, on limited platforms, and with significant memory constraints (Tan & Varga, 2008). Of course, ongoing advances in semiconductor technologies mean that such constraints will naturally become less significant over time.

In fact, increased computing resources coupled with more sophisticated software methods may be expected to narrow the performance differential between embedded and server-based recognition applications: the border between applications realized by these techniques will narrow, allowing for advanced features such as natural language understanding to become possible in an embedded context rather than simple command-and-control systems. At this point there will no longer be significant technological barriers to use of embedded systems to create a smart VI-enabled home.

However at present, embedded devices typically have relatively slow memory access, and a scarcity of system resources, so it is necessary to employ a fast and lightweight speech recognition engine in such contexts. Several such embedded ASR systems have been introduced in (Hataoka et al., 2002), (Levy et al., 2004), and (Phadke et al., 2004) for sophisticated human computer interfaces within car information systems, cellular phones, and interaction device for physically handicapped persons (and other embedded applications) respectively.

It is also possible to perform speech recognition in smart homes by utilising a centralised server which performs the processing, connected to a set of microphones and loudspeakers scattered throughout a house: this requires significantly greater communications bandwidth than a distributed system (since there may be arrays of several microphones in each location, each with 16 bit sample depth and perhaps 20kHz sampling rate), introduces communications delays, but allows the ASR engine to operate on a faster computer with fewer memory constraints.

As the capabilities of embedded systems continue to improve, the argument for a centralised solution will weaken. We confine the discussion here to a set of distributed embedded systems scattered throughout a smart home, each capable of performing speech recognition, and VI. Low-bandwidth communications between devices in such a scenario to allow co-operative ASR (or CPU cycle-sharing) is an ongoing research theme of the authors, but will not affect the basic conclusions at this stage.

In the next section, the open source Sphinx is described as a reasonable choice among existing ASRs for smart home services. We will explain why Sphinx is suitable for utilisation in smart homes as a VI core through examining its capabilities in an embedded speech recognition context.

8. Sphinx as an ASR for smart homes

Among many automatic speech recognizers available for different applications with various features, the open source Sphinx recognizer is an excellent example of a flexible modern speech recognition system. Sphinx, originally developed at Carnegie Mellon University in the USA, provides and integrates several capabilities that allow it to be adapted for a wide range of different speech recognition applications.

At one extreme, it can be used for single word recognition, or expanded at the other extreme to large vocabularies containing tens of thousands of words. In terms of resource constraints, it can run on anything from a tiny embedded system (PocketSphinx) to a large and powerful server (which could run the Java language version Sphinx-4). Sphinx is regularly updated and evaluated within the speech recognition research field.

Sphinx, in common with most current ASR implementations, relies upon Hidden Markov Modelling to match speech features to stored patterns (Lee, 1989). It is highly configurable and incredibly flexible – the required features used can be selected as required.

Sphinx2, the decoding engine for Sphinx II, can be a good choice for smart home services, provided several appropriate model files and databases are used. These are classified into three categories:

- a. Pronunciation lexicon/dictionary defining words of current interest, and a phonemic pronunciation for each.
- b. Acoustic models based on Hidden Markov Models (HMM) for base phones and triphones. Sphinx2 uses both semi-continuous and continuous density acoustic models which are typically generated by the Sphinx acoustic model trainer.
- c. A predetermined language model accepting two flavours of language: either the finite state graph (FSG) and N-gram models (where N is either two or three).

Apart from ordinary words, noise or filler words can be specified for a particular application by placing them in a corresponding dictionary. The N-gram language model additionally includes begin-sentence and end-sentence symbols, denoted <S> and </S>, normally representing silence. These can be used in continuous speech applications for quiet homes, but may need to be augmented with predetermined start/stop attention phrases.

The core speech decoder operates on finite-length segments of speech or utterance, one utterance at a time. An utterance can be up to one minute long, but in practice most applications handle sentences or phrases which are much shorter than this. For real-time use, processing must be continuous, with a response latency that is not excessive. Response delays of a second or more may well lead to user annoyance.

As mentioned in section 3, smart home services are a mixture of small (for command-and-control applications) and large (for email dictation and similar applications) continuous vocabulary speech systems; thus, we need an ASR which supports both modes. As a comparison, the concept is similar to (Nakano et al., 2006) in which the Honda ASIMO humanoid robot has two dialogue strategies: a) task-oriented dialogues which utilize the outputs of a small vocabulary speech recognizer, and b) non-task-oriented dialogues which utilize the outputs of a large vocabulary speech recognizer.

The major difference between Sphinx and this approach occurs during the implementation phase where ASIMO deploys two different ASR engines (Julian for small vocabulary and Julius for large one). This differs to the authors Sphinx-based system which proposed a single recognition engine that not only caters for the needs of both tasks, but has a continuously variable vocabulary instead of two extremes as in the ASIMO case. This

therefore allows a continuum of dialogue complexities to suit the changing needs of the vocal human-computer interaction. The particular vocabulary in use at any one time would depend upon the current position in the grammar syntax tree.

As a noticeable choice in embedded applications necessary for smart homes, Sphinx II is available in an embedded version called PocketSphinx. Sphinx II was the baseline system for creating PocketSphinx because it is faster than other recognizers currently available in the Sphinx family (Huggins-Daines et al., 2006). The developers claim PocketSphinx is able to address several technical challenges in deployment of speech applications on embedded devices. These challenges include computational requirements of continuous speech recognition for a medium to large vocabulary scenario, the need to minimize the size and power consumption for embedded devices which imposes further restrictions on capabilities and so on (Huggins-Daines et al., 2006).

Actually, PocketSphinx, by creating a four-layer framework including: frame layer, Gaussian mixture model (GMM) layer, Gaussian layer, and component layer, allows for straightforward categorisation of different speed-up techniques based upon the layer(s) within which they operate.

9. Audio aspects

As mentioned in section 1, smart home VI provides a good implementation target for practical ASR: the set of users is small and can be predetermined (especially pre-trained, and thus switched-speaker-dependent ASR becomes possible), physical locations are well-defined, the command set and grammar can be constrained, and many noise sources are already under the control of (or monitored by) a home control system.

In terms of the user set, for a family home, each member would separately train the system to accommodate their voices. A speaker recognition system could then detect the speech of each user and switch the appropriate acoustic models into Sphinx. It would be reasonable for such a system to be usable only by a small group of people.

Physical locations – the rooms in the house – will have relatively constant acoustic characteristics, and thus those characteristics that can be catered for by audio pre-processing. Major sources of acoustic noise, such as home theatre, audio entertainment systems, games consoles and so on, would likely be under the control of the VI system (or electronically connected to them) so that methods such as spectral subtraction (Boll, 1979) would perform well, having advanced knowledge of the interfering noise.

It would also be entirely acceptable for a VI system, when being required to perform a more difficult recognition task, such as LVCSR for email dictation, to automatically reduce the audio volume of currently operating entertainment devices.

Suitable noise reduction techniques for a smart home VI system may include methods such as adaptive noise cancellation (ANC) (Hataoka et al., 1998) or spectral subtraction which have been optimized for embedded use (Hataoka et al., 2002).

The largest difference between a smart home ASR deployment and one of the current computer-based or telephone-based dictation systems is microphone placement (McLoughlin, 2009): in the latter, headset or handset microphones are used which are close to the speakers mouth. A smart home system able to respond to queries anywhere within a room in the house would have a much harder recognition task to perform. Microphone arrays, steered by phase adjustments, are able to 'focus' the microphone on a speakers mouth (Dorf, 2006), in some cases, and with some success.

However more preferable is a method of encouraging users to direct their own speech in the same way that they do when interacting with other humans: they turn to face them, or at least move or lean closer. This behaviour can be encouraged in a smart home by providing a focus for the users. This might take the form of a robot head/face, which has an added advantage of being capable of providing expressions – a great assistance during a dialogue when, for example, lack of understanding can be communicated back to a user non-verbally. This research is currently almost exclusively the domain of advanced Japanese researchers: see for example (Nakano et al., 2006).

A reasonable alternative is the use of a mobile device, carried by a user, which they can speak into (Prior, 2008). This significantly simplifies the required audio processing, at the expense of requiring the user to carry such a device.

10. Usability criteria

It is sensible to develop smart home VI usability criteria. Users of current ASR systems may well appreciate the frustrations of mis-tuned and under-performing systems where backtracking and corrections require more effort and time than does the process of input itself. Many give up, preferring to switch back to a clumsy, but reliable, keyboard.

Despite the theoretical reasons for adopting a VI system to free users from the constraint of KMM, most users have in fact grown up with such constraints and are not unhappy with them. By contrast, users of current ASR systems tend to experience mostly frustration in their interactions – a major reason why LVCSR is not particularly popular today. It is therefore only when viable alternatives to the KMM have been demonstrated in niche areas that the general public will adopt a new perspective on the use of computer technology without touch and vision-based user interfaces. Such a niche area is likely to be within the smart home context, where significant advantages exist for ASR: a limited set of users, relatively constant acoustic characteristics, constraints upon the tasks to be performed and so on.

The assumptions for a smart home are that training is performed in advance, the major sources of acoustic interference (such as infotainment and gaming systems) are under control or at least linked in to the smart home electronics, the VI operational syntax is self-contained, limited and known to the user, and that the user has had time to become familiar with the system.

Major performance criteria include the percentage of tasks which are completed without secondary user intervention (and then the degree of intervention required for those tasks which are not completed straight off). For use by the general public, a useful performance measure may well be how often the user must resort to a KMM solution for performing VI-oriented tasks. However, as with all technology deployments to the general public, a final successful verdict can only be pronounced when sales figures begin to indicate mass-marked adoption of such technology.

11. Conclusion

The major components of a smart home ASR system currently exist within the speech recognition research community, as the evolutionary result of half a century of applied and academic research. The command-and-control application of appliances and devices within the home, in particular the constrained grammar syntax, allows a recognizer such as Sphinx

to operate with high levels of accuracy. Results are presented here which relate accuracy to vocabulary size, and associate metrics for reducing vocabulary (and thus maximising accuracy) through the use of restricted grammars for specialised applications.

Audio aspects related to the smart home, and the use of LVCSR for multi-user dictation tasks are currently major research thrusts, as is the adaption of ASR systems for use in embedded devices. The application of speech recognition for performing WWW queries is probably particularly important for the adoption of such systems within a usable smart home context, and this work is ongoing, and likely to be greatly assisted if current research efforts towards a semantic web will impact the WWW as a whole.

The future of ASR within smart homes will be assured first by the creation of niche applications which deliver to users in a friendly and capable fashion. That the technology largely exists has been demonstrated here, although there is still some way to go before such technology will be adopted by the general public.

12. References

- Boll, S. F. (1979). "Suppression of acoustic noise in speech using spectral subtraction", *IEEE Transactions on Signal Processing*, Vol. 27, No. 2, pp. 113-120.
- Chevalier, H.; Ingold, C.; Kunz, C.; Moore, C.; Roven, C.; Yamron, J.; Baker B.; Bamberg, P.; Bridle, S.; Bruce, T.; Weader, A. (1996) "Large-vocabulary speech recognition in specialized domains", *Proc. ICASSP*, Vol. 1 pp. 217-220.
- Chow, Y.L.; Dunham, M.O.; Kimball, O. A.; Krasner, M. A.; Kubala, G. F ; Makhoul, J.; Roucos, S. ; Schwartz, R. M. (1987). "BBYLOS: The BBN continuous speech recognition system," *Proc. ICASSP.*, pp.89-92.
- Davis, K. H.; Biddulph, R.; Balashek, S. (1952). "Automatic recognition of spoken digits", *J. Acoust. Soc. Am.*, Vol 24, No. 6.
- Dorf, C. (2006). *Circuits, Signals, and Speech And Image Processing*, CRC Press.
- ECHONET Consortium (2008). *Energy Conservation and Homecare Network*, www.echonet.gr.jp, last accessed July 2008.
- Fry, D. B.; Denes, P. (1959). "The design and operation of the mechanical speech recognizer at University College London", *J. British Inst. Radio Engr.*, Vol. 19, No. 4, pp. 211-229.
- Furui, S. (2001). "Toward flexible speech recognition-recent progress at Tokyo Institute of Technology", *Canadian Conference on Electrical and Computer Engineering*, Vol. 1, pp. 631-636.
- Glass, J.; Weinstein, E. (2001). "SpeechBuilder: Facilitating Spoken Dialogue System Development", *7th European Conf. on Speech Communication and Technology*, Aalborg Denmark, pp. 1335-1338.
- Gorin, A. L.; Parker, B. A.; Sachs, R. M. and Wilpon, J. G. (1996). "How May I Help You?", *Proc. Interactive Voice Technology for Telecommunications Applications (IVTTA)*, pp. 57-60.
- Hataoka, N.; Kokubo, K.; Obuchi, Y.; Amano, A. (1998). "Development of robust speech recognition middleware on microprocessor", *Proc. ICASSP*, May, Vol. 2, pp. 837-840.
- Hataoka, N.; Kokubo, K.; Obuchi, Y.; Amano, A. (2002). "Compact and robust speech recognition for embedded use on microprocessors", *IEEE Workshop on Multimedia Signal Processing*, pp. 288-291.

- Huggins-Daines, D.; Kumar, M.; Chan, A.; Black, A. W.; Ravishankar, M.; Rudnicky, A. I. (2006). "PocketSphinx: a free, real-time continuous speech recognition system for hand-held devices", Proc. ICASSP, Toulouse.
- Itakura, F. (1975). "Minimum prediction residual applied to speech recognition", IEEE Transactions on Acoustics, Speech, Signal Processing, pp.67-72.
- Kamm, C. A.; Yang, K.M.; Shamieh, C. R.; Singhal, S. (1994). "Speech recognition issues for directory assistance applications", 2nd IEEE Workshop on Interactive Voice Technology for Telecommunications Applications IVTTA94, May, pp. 15-19, Kyoto.
- Kango, R.; Moore, R.; Pu, J. (2002). "Networked smart home appliances - enabling real ubiquitous culture", Proceedings of 5th International Workshop on Networked Appliances, Liverpool.
- Kango, R.; Pu, J.; Moore, R. (2002b). "Smart appliances of the future - delivering enhanced product life cycles", The 8th Mechatronics International Forum Conference, University of Twente, Netherlands.
- Kryter, K. D. (1995). *The Handbook of Hearing and the Effects of Noise*, Academic Press.
- Lahrman, A. (1998). "Smart domestic appliances through innovations", 6th International Conference on Microsystems, Potsdam, WE-Verlag, Berlin.
- Lee, K. F. (1989). *Automatic Speech Recognition: The Development of the Sphinx System*, Kluwer Academic Publishers.
- Lee, K. F. ; Hon, H. W.; Reddy, D. R. (1990). "An overview of the Sphinx speech recognition system", IEEE Transactions on Acoustics, Speech, Signal Processing, vol.38(1), Jan, pp. 35-45.
- Lee, C. H.; Rabiner, L. R.; Peraccini, R.; Wilpon, J. G. (1990). "Acoustic modeling for large vocabulary speech recognition", *Computer Speech and Language*.
- Levy, C.; Linares, G.; Nocera, P.; Bonastre, J. (2004). "Reducing computational and memory cost for cellular phone embedded speech recognition system", Proc. ICASSP, Vol. 5, pp. V309-312, May.
- Martin, T. B.; Nelson, A. L.; Zadell, H. J. (1964). "Speech recognition by feature abstraction techniques", Tech. Report AL-TDR-64-176, Air Force Avionics Lab.
- McLoughlin, I.; Sharifzadeh, H. R. (2007). "Speech recognition engine adaptations for smart home dialogues", 6th Int. Conference on Information, Communications and Signal Processing, Singapore, December.
- McLoughlin, I. (2009). *Applied Speech and Audio*, Cambridge University Press, Jan.
- McTear, M. F. (2004). *Spoken Dialogue Technology: Toward The Conversational User Interface*, Springer Publications.
- Miller, G. A.; Heise, G. A.; Lichten, W. (1951). "The intelligibility of speech as a function of the context of the test materials", *Exp. Psychol.* Vol. 41, pp. 329-335.
- Minker, W.; Bannacef, S. (2004). *Speech and Human-Machine Dialog*, Kluwer Academic Publishers.
- Nakano, M.; Hoshino, A.; Takeuchi, J.; Hasegawa, Y.; Torii, T.; Nakadai, K.; Kato, K.; Tsujino, H. (2006). "A robot that can engage in both task-oriented and non-task-oriented dialogues", 6th IEEE-RAS International Conference on Humanoid Robots, pp. 404-411, December.
- Paul, D. B. (1989). "The Lincoln robust continuous speech recognizer," Proc. of ICASSP, vol.1, pp. 449-452.

- Phadke, S.; Limaye, R.; Verma, S.; Subramanian, K. (2004). "On design and implementation of an embedded automatic speech recognition system", 17th International Conference on VLSI Design, pp. 127-132.
- Prior, S. (2008). "SmartHome system", <http://smarthome.geekster.com>, last accessed July 2008.
- Rabiner, L. R.; Levinson, S. E.; Rosenberg, A. E.; Wilpon, J. G. (1979). "Speaker independent recognition of isolated words using clustering techniques", IEEE Transactions on Acoustics, Speech, Signal Processing, August.
- Rabiner, L. R. (1989). "A tutorial on hidden markov models and selected applications in speech recognition", Proc. IEEE, pp. 257-286, February.
- Rabiner, L. R. (1994). "Applications of voice processing to telecommunications", In proceedings of the IEEE, Vol. 82, No. 2, pp. 199-228, February.
- Ravishankar, M. K. (1996). "Efficient algorithms for speech recognition", Ph.D thesis, Carnegie Mellon University, May.
- Roy, D. (1999). "Networks for homes", IEEE Spectrum, December, vol. 36(12), pp. 26-33.
- Sakoe, H.; Chiba, S. (1978). "Dynamic programming algorithm optimization for spoken word recognition", IEEE Transactions on Acoustics, Speech, Signal Processing, February, vol.26(1), pp. 43-49.
- Sun, H.; Shue, L.; Chen, J. (2004). "Investigations into the relationship between measurable speech quality and speech recognition rate for telephony speech", Proc. ICASSP, May, Vol. 1, pp.1.865-1.868.
- Tan, Z. H.; Varga, I. (2008). Automatic Speech Recognition on Mobile Devices and over Communication Networks, Springer Publications, pp. 1-23.
- Tanenbaum, A. (1996). Computer Networks, 3rd ed. Upper Saddle River, N.J. London, Prentice Hall.
- Velichko, V. M.; Zagoruyko, N. G. (1970). "Automatic recognition of 200 words", International Journal of Man-Machine Studies, June, Vol.2, pp. 223-234.
- Wang, Y. M.; Russell, W.; Arora, A.; Jagannathan, R. K. Xu, J. (2000). "Towards dependable home networking: an experience report", Proceedings of the International Conference on Dependable Systems and Networks, p.43.
- Weibel, A.; Hanazawa, T.; Hinton, G.; Shikano, K.; Lang, K. (1989). "Phoneme recognition using time-delay neural networks", IEEE Transactions on Acoustics, Speech, Signal Processing, March, Vol.37(3), pp. 328-339.
- Wikipedia, (2008). http://en.wikipedia.org/wiki/Semantic_web, last accessed July 2008.
- Zue, V.; Glass, J.; Phillips, M.; Seneff, S. (1989). "The MIT summit speech recognition system: a progress report", Proceedings of DARPA Speech and Natural Language Workshop, February, pp. 179-189.