Keller, E. & Caelen, J. (1994). Introductions to Sections 1, 2, & 3 Background, State of the Art, & Challenges. In E. Keller (ed.), *Fundamentals of Speech Synthesis and Speech Recognition: Basic Concepts, State of the Art, and Future Challenges* (pp. 1-4, 63-68, 171-183). Chichester: John Wiley.

SECTION 1

BACKGROUND

SECTION 1: BACKGROUND

E. Keller and Jean Caelen
Introduction to Section 1: Background

E. Keller
Fundamentals of Phonetic Science

S. Werner
Prosodic Aspects of Speech

B. Zellner
Pauses and the Temporal Structure of Speech

Introduction to Section 1: Background

Eric Keller[1] and Jean Caelen[2]

[1] Université de Lausanne, LAUSANNE, Switzerland
[2] Institut de la Communication Parlée, GRENOBLE, France

The essential representation of speech on the computer is the acoustic signal (Figure 1). During speech input, acoustic waveforms or impulses are captured by a microphone, and are translated into long strings of numbers, i.e., digital acoustic signals, before they are interpreted by speech recognition algorithms[1]. During speech output, the final, synthetic form is nearly always an analogous acoustic signal, which now traverses the inverse path between numeric form and the acoustic waveform issuing from a loudspeaker or earphones. The I/O processes mediating between the acoustic waveform and the corresponding digital signal are handled in fairly standard fashion, and are thus not of interest here. The crucial relationship that a computer has to

[1] Please note the use of the term "speech recognition", instead of the popular term "voice recognition", throughout this volume. The difference is semantically relevant. The recognition of the voice of a particular speaker, as suggested by the term "voice recognition", is used in applications that determine access to information or to locales on the basis of the identification of voice characteristics of a speaker. As Chapter 7 by Gérard Chollet indicates, "voice recognition" or "speaker recognition" is a special area of investigation, quite separate in its aims, objectives and working principles from the area of speech recognition proper.

deal with in speech synthesis and speech recognition is the higher-level relationship between the acoustic signal and a chain of language symbols.

The first few chapters of this book explain the main characteristics of this relationship. In the first chapter by Keller, the emphasis is on the basic nature of the link between a given speech sound (a "phoneme" or a "speech segment"), its articulatory realisation and its acoustic waveform. This chapter illustrates therefore the segmental aspect of speech and provides an explanation of the basic concepts of speech analysis.

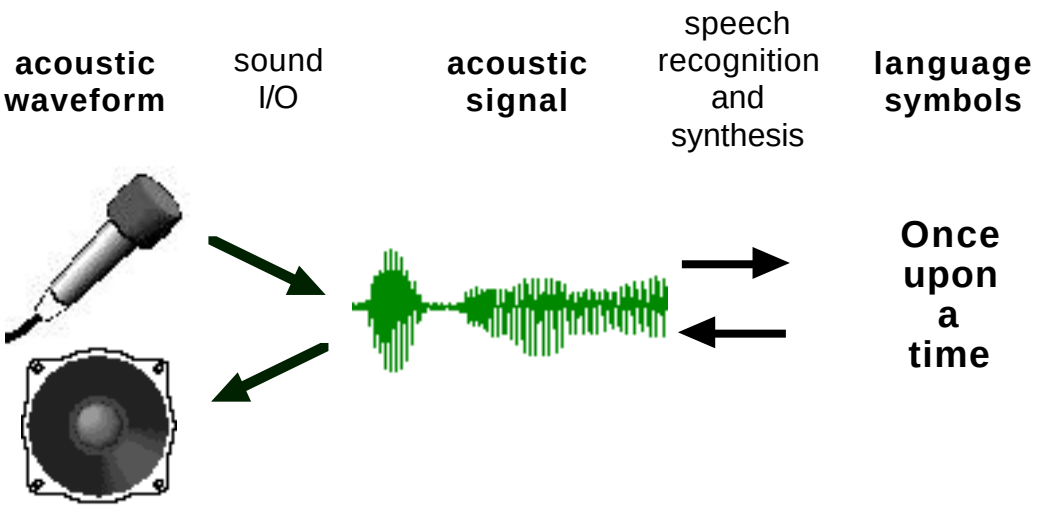

Figure 1. The position of speech recognition and speech synthesis in the relationship between sound input/output (I/O), acoustic signals and language symbols.

In the chapter by Werner and Keller, relationships involving longer-term parameters, known as "supra-segmental" or "prosodic" parameters, are explained. Common supra-segmental parameters are intonation, rhythm and phrasal timing. Prosodic parameters are used to distinguish between statements ("declarative sentences"), questions ("interrogative sentences"), and commands ("imperative sentences"). They are also employed to clarify the sentence structure and to indicate emphasis.

The third chapter by Zellner addresses a related set of concepts in the organisation of speech, temporal structure and pausing. The importance of this area arises from the very nature of the link between the acoustic signal and the various elements of the language chain. Much as in music, the key element of successful synthesis is the correctly-timed

orchestration of all components. Synthesised speech with bad timing is just as difficult to understand as speech made up of the wrong speech sounds. If computer speech is to sound anything like human speech, if it is to be comfortable and pleasant to listen to, it will therefore have to implement pauses very much as human speakers do. Consequently a sophisticated understanding of speech timing is required for the implemention of satisfactory synthesis models.

Even speech recognition — which so far has ignored timing and pausing issues — may benefit from this type of knowledge. Zellner indicates that speech employs pauses to structure the various chunks of meaning in a message. Pauses may ultimately be just as important as segmental information for deriving meaning from connected speech.

SECTION 2

STATE OF THE ART

SECTION 2: STATE OF THE ART

E. Keller and Jean Caelen
Introduction to Section 2: State of the Art

P. Bhaskararao
Subphonemic Segment Inventories for Concatenative Speech Synthesis

B. Pfister and C. Traber
Text to Speech Synthesis: An Introduction and a Case Study

T. Styger and E. Keller
Formant Synthesis

G. Chollet
Automatic Speech Recognition: State of the Art, Current Issues and Perspectives

K. Torkkola
StochasticModels and Artificial Neural Networks for Automatic Speech Recognition

Introduction to Section 2: State of the Art

Eric Keller[1] and Jean Caelen[2]

[1] Université de Lausanne, LAUSANNE, Switzerland
[2] Institut de la Communication Parlée, GRENOBLE, France

In speech synthesis and speech recognition, the "state of the art" is a moving target. For this reason, no attempt was made to provide extensive descriptions of existing synthesis and recognition systems in this book. Instead, it was decided to concentrate on the fundamental techniques of the field. While marketed systems evolve rapidly, and are often shrouded in commercial secrecy, the fundamental techniques they employ remain quite stable. The chapters of this section introduce these techniques, and they provide a great many useful references to further readings in the field.

The section begins with a number of chapters on speech synthesis. Synthesised speech is in essence "re-composed speech", i.e. speech reconstituted in one way or another from elements or portions of previously recorded speech. One of the crucial issues in synthesis concerns the elements that go into the reconstitution.

At the most banal level, one can simply record whole sentences and play them back as required. Indeed, for many common applications, this remains the simplest and most efficient solution (e.g. "4th floor Household appliances"). At the next level, part of the recording is stable, and part is generated from recordings of single words — again a familiar solution for information systems. For example, telephone companies often provide phone numbers according to the pattern: "The number is: 9-8-7, 3-5-4-6". The initial portion of the message is

recorded, and the numbers are generated from individual recordings. In order to facilitate comprehension, the number preceding the comma is chosen from a special set of numbers having "pause characteristics", i.e. lengthened final syllables and special intonation contours.

However, such solutions tend to be of marginal interest. The real challenge lies in reconstructing almost any speech from much smaller segments or even, from a set of abstract elements. The three initial chapters of this section discuss this issue from different perspectives.

In the first chapter, Bhaskararao describes experiments performed with minimalist or near-minimalist systems. In this case, exceedingly small segments of speech are stored and concatenated as required. This approach is appropriate to cases where computing power is weak and available storage is minimal (Bhaskararao's own segment inventory for Indian languages is just 1.6 Mb in uncompressed form). However, for reasons indicated in the previous section, coarticulatory information — and with it, a great deal of naturalness of speech — is all but lost in such minimalist systems. In the second part of his chapter, Bhaskararao explains how coarticulatory information is best incorporated into somewhat more powerful systems, while respecting, as best as possible, the combinatorial possibilities of various types of languages.

Pfister and Traber present the various choices that go into designing the synthesis architecture for a mature concatenative system, capable of dealing with free connected text. From the perspective of their own, highly respected SVOX system, they show that beyond the creation of a satisfactory concatenative segment inventory lies considerable additional footwork associated with developing the grammatical underpinnings of a synthesis system. Grammatical knowledge ("syntax" and "morphology"[2]) is required to disambiguate so-called homographs and to provide realistic-sounding prosodic

---

[2] *Syntax: The systematic order that language elements follow in an utterance to convey a particular grammatical sense. Example: The difference between normal (declarative) and inverted (interrogative) word order, as "are you" vs. "you are" in "are you/you are in a good mood". Morphology: Modifications of a base form designed to a convey specific grammatical sense. Example: "see" — "sees" — "seeing" or "rose" — "roses".*

information. A common example of an English homograph is the word "record" which has two different pronunciations depending on its grammatical class and function (e.g. "she rec'ords 'records"). Grammatical knowledge permits to identify which form of the word is present, and thus permits the selection of the correct pronunciation.

Styger and Keller present a totally different approach to the problem of speech synthesis. Instead of attempting to "glue together" pieces of pre-recorded speech, relatively abstract parameters, such as frication, formants and amplitudes, are stored away and recomposed as required. This approach is called "formant synthesis", and over short stretches of speech, it has been shown to be capable of generating utterances that are difficult to distinguish from those produced by human speakers.

The reason behind this outstanding performance again has much to do with coarticulation. Coarticulatory effects can extend over considerable periods of time, and often span more time than even the longest stored segments of a concatenative system. So even very good and extensive inventories of speech segments cannot account for all coarticulatory effects that are possible in a language. Advanced formant synthesis systems, on the other hand, contain their own rules for simulating coarticulatory effects, which are then applied to the calculation of the right combination of formants.

Another improvement comes from the generation of the intonational contour. Segment-based systems generally perform considerable signal-analytic sleight-of-hand to stretch or squeeze the fundamental frequency period of a speech signal into a shape that produces the auditory effects of higher or lower voice pitch. Although modern techniques of this sort can be very good, they do leave a distinct auditory trace. Formant synthesis, on the other hand, generates pitch from first principles, and thus has the potential of producing much cleaner speech waveform.

However, the great promise of this approach is also its greatest pitfall. Typical formant synthesis systems manipulate arrays of between 40 and 60 parameters, and generate a new set of parameters every 5-10 ms. Opinions are still sharply

[3] They apply a so-called PSOLA, a "pitch-synchronous overlap and add" operation to the signal (see Pfister and Traber's chapter).

divided on whether this is a practicable task. John Local (York University, UK, see Chapter 12) has shown that outstanding results are possible in the generation of any two-syllable English word by the application of a few hundred, well-designed rules. However some other authors remain exceedingly sceptical of the general viability and commercial applicability of this approach.

The final two chapters in this section by Chollet and Torkkola deal with speech recognition. Gérard Chollet has taken upon himself the considerable task of providing a first introduction — in just a few pages — to the enormous field of speech and speaker recognition. The essential question in recognition concerns the ultimate application. Command recognition to control a computer, a chirurgical instrument or aircraft operation is rather different from the recognition of connected speech in interactive information systems. Recognition without any or with very limited training involves quite different algorithms and currently makes much more limited promises, than recognition after a more or less considerable training period. Finally, speaker recognition is quite different from speech recognition. Required computing power and algorithms, as well as the system's ultimate constraints are directly related to the specific purpose that the recognition capacity is supposed to satisfy.

Torkkola's chapter links up directly with these issues by discussing the two main computational approaches to recognition. In contrast to synthesis where phonetic and linguistic considerations are in the foreground, statistical and still to a lesser degree, neuro-computational algorithms are in the foreground in current-day speech recognition. Torkkola explains in some detail the advantages of both approaches. Statistical (or "Markovian") techniques predominate in current applications, but for a number of relatively technical reasons, neuro-mimetic approaches show considerable promise of being able to surpass statistical approaches. Their ascendancy in desktop computing has recently been reinforced by the mass production of affordable neurocomputing chips (Baran, 1994). It is likely that this type of computing capacity will soon be applied to speech recognition where it will no doubt have a massive impact.

Reference

Baran, N. (1994). Neural networks: Intel and Nestor to commercialize neural-net chip.BYTE, 19 (March),32.

SECTION 3

CHALLENGES

SECTION 3: CHALLENGES

E. Keller and Jean Caelen
Introduction to Section 3: Challenges

L.-J. Boë, J.-L. Schwartz and N. Vallée
The Prediction of Vowel Systems: Perceptual Contrast and Stability

B. Gabioud
Articulatory Models in Speech Synthesis

P. Perrier and D. Ostry
Dynamic Modelling and Control of Speech Articulators: Application to Vowel Reduction

J. Local
Phonological Structure, Parametric Phonetic Interpretation and Natural-sounding Speech Synthesis

G. Caelen
Semantic and Pragmatic Prediction of Prosodic Structures

M. Cooke and G. Brown
Separating Simultaneous Sound Sources: Issues, Challenges and Models

Q. Summerfield and J. Culling
Auditory Computations that Separate Speech from Competing Sounds:
A Comparison of Monaural and Binaural Processes

J. Caelen
Multimodal Human-Computer Interface

Introduction to Section 3: Challenges

Eric Keller[1] and Jean Caelen[2]

[1] Université de Lausanne, LAUSANNE, Switzerland
[2] Institut de la Communication Parlée, GRENOBLE, France

The "Challenges" section of this book occupies about half of the chapters. The reader is warned that there is a substantial jump in difficulty as one moves from the first to the second half of this volume. If the subsequent chapters define some "future challenges" for our field in general, they may also represent a "more immediate comprehension challenge" for the new reader.

For this reason, we begin the section with a somewhat more detailed discussion of the central issue which inspired the assembling of the "Challenges" chapters in the first place. It is the question of how much knowledge about human speech processing is required for developing speech synthesis and speech recognition devices. Many of the arguments in the first half of this book have already demonstrated that speech synthesis cannot operate without direct reference to the human model. Without being a carbon copy of human speech production, synthesis must necessarily be structured along parameters and principles that are similar to those that govern human speech. However, is the same true of speech recognition? Many observers resolutely maintain that that is not so.

In automatic speech recognition, a difference can be made between systems based on artificial intelligence (AI), and those that emerge from various techniques of pattern recognition

and thus proceed by the identification of prototypes (DTW: "Dynamic Time Warping", HMM: "Hidden Markov Model", NN: "Neural" or "Neuromimetic Network"). Overall, the latter systems (e.g. the Markovian models) have shown better performance than AI-type systems, without being by any means inspired or necessarily similar in operation to human perceptual and cognitive processes. Indeed, judging exclusively by results obtained, or by systems in current use, it would seem that reference to the human cognitive model is not at all required for good performance in speech recognition. As a consequence, much cognitively-, linguistically- or AI-oriented research in speech recognition has been totally abandoned over the last few years.

We would argue that this is a regrettable and short-sighted policy. It would seem to us that over the long term, the systematic neglect of issues surrounding human communicative, cognitive and linguistic behaviour is likely to lead to substantial confusion and to major limitations in recognition systems. Systems that recognise by a set of "magic black boxes" (e.g. large arrays of markovian chains or neuromimetic pattern recognisers, dealing in global fashion with speech input) cannot be expected to produce more than the correct recognition of pre-stored speech sequences and the rejection of non-stored patterns. Unless a system is designed to learn specific types of cognitive or linguistic information, it cannot be expected to generalise, or to "learn" parameters in terms of familiar cognitive and linguistic categories. Consequently, when such systems fail, debugging is severely hampered by a lack of evident coherence, and by the lack of explicitness of system parameters. Much like reptilians that never evolved beyond a certain highly efficient pattern recognition, pattern-driven speech recognition cannot be expected to evolve much beyond a fairly focused, and thus admittedly highly efficient, recognition performance.

This is not to say that human computation cannot be efficient as well. Current knowledge on the computations that humans perform in order to understand complex communicative interactions, for example, provide ample reason to argue that human computations, while very complex, can also be highly efficient (for some appreciation of the details of this argument, see Chapter 16). The supposed inadequacies of human computation is therefore insufficient reason for

rejecting all anthropomorphic computer models of speech recognition. Quite the contrary. Studies of human communicative behaviour suggest that human speaker-listeners operate with exceeding efficiency, and take into account complex interactions between supporting forms of evidence, in order to recognise and interpret spoken information. The issue is not so much one of efficiency of computation, as it is an issue of knowing exactly what human computations are performed in a given situation to arrive at a specific recognition or interpretation performance.

To further cement this argument, let us examine the various aspects of human recognition performance. This may guide us in establishing which human-type computations might be of interest when creating similar computer-based algorithms.

1. Human perception is relatively robust in the parameter extraction of relevant acoustic data from noisy acoustic input (cf. Chapters 14 and 15).

2. Human perception is fairly successful at calculating distinctive features (be they articulatorily or acoustically based) for the discrimination of specific sounds or sound patterns (see Chapter 9).

3. Human listeners show good performance at the identification (sometimes "the active reconstruction") of those phonological and lexical events that are most helpful in understanding a given passage — and that despite the fact that such events may be coarticulatorily and prosodically encoded in literally thousands of different ways (see also Chapter 9).

4. Human listeners appear to maintain sets of local identification hypotheses, which indicates that they are capable of dealing in an efficient manner with uncertain and incompletely-decoded material.

5. Humans can exercise complex reasoning in evaluating, completing or rejecting such partial hypotheses on the basis of multiple sources of heterogeneous knowledge (acoustic, phonetic, lexical, syntactic, semantic or pragmatic knowledge).

6. Humans can apply sophisticated decision-making and control strategies in order to coordinate processing, and to resolve coreferential language phenomena such as anaphoras, deictics, hesitations, error corrections, ellipses, incomplete sentences, etc. Even more sophisticated coordinate processing

occurs when multi-modal input is analysed (see Chapter 16).

7. Humans are well-equipped to adapt their perceptual mechanisms to novel situations and to new speakers, or to new and complex tasks. Also, humans show exceptional ability to learn new vocabulary and even new language structures throughout their lives.

8. Current artificial recognition systems are comparatively ill-prepared and easily derouted in face of a series of higher-level tasks. For example, they cannot "guess" at the meaning of a new, unfamiliar word the way human language users can, they cannot coalesce information from multiple sources (e.g. when several speakers talk at the same time), nor can they compensate for erroneous pronunciations, or adjust to a strong foreign accent. The best current systems (the statistically-based systems) are not advanced enough to be able to draw inferences from and to improve upon their own performance; both of these are abilities in which humans excel. Consider also that humans can reason about events that have never even occurred, and which are therefore (by definition) beyond the grasp of statistical systems.

Reflections such as these amply motivate the position we've taken in this volume. In our perspective, development on synthesis as well as recognition devices can profit from looking at human speech performance. However, this position must be understood correctly. We do not propose that machines should be constructed exactly like humans. As proponents of the non-cognitive approach like to point out, when humans learned to fly, the winning design was not an articulated, highly flexible, bird-like wing, but it was a relatively simple, fixed-wing design. For a variety of excellent reasons, recognition devices built for machines will necessarily operate differently than those of humans.

However, consider also that it wasn't by avoiding the force of lift and the law of gravity that flight was mastered. In terms of the flight analogy, the current status of automatic speech recognition bears greater resemblance to the achievements of the Wright brothers than to current fixed- or rotary-wing flight. The "thing" flies, often just barely, but a great deal more theory needs to be understood and intelligently applied to operational systems to insure "secure flight" under the most exacting conditions. With this in mind, let us examine in some more detail a number of current approaches to speech

recognition.

Which Speech Recognition System?

Outside of the statistical systems, the principles of which are explained in Torkkola's chapter (Chapter 8), the majority of current knowledge-oriented systems are constituted of a community of "experts" (or "agents") that exchange information in a proposal/verification mode. For example, there might exist an "expert" on fundamental frequency whose job is to identify different prosodic structures in the input stream. These recognition systems are mainly distinguished by the strategy that they employ, more than by the knowledge sources that they use. In fact, all of these systems manipulate knowledge about roughly the same acoustic, phonetic, lexical, prosodic, syntactic and semantic information.

In such systems, the "experts" — each related to one or several knowlege sources — operate either in coordinated fashion under the direction of an external supervisor, in autonomous fashion, or in distributed fashion. In the case of coordinated operation under a supervisor, the recognition strategy is determined by a centralised process which activates the experts according to a hierarchical (Sacerdotti, 1977) or an "opportunist"[4] plan (Hayes-Roth and Hayes-Roth, 1979). Several layers of experts, and thus several supervisors, are also possible (Gong and Haton, 1987; Sabah, 1988, 1990).

In the cases of autonomous and distributed systems, the agents cooperate actively to resolve the problems with which they are confronted. They not only communicate to exchange data and knowledge which they require to solve a problem, but they also maintain control information about their own status. In these latter systems, a level of metacommunication is thus required where agents must inform each other about their internal status. We will examine the major types of these autonomous or distributed models next.

The Multi-expert Models

[4] *In the sense of "exploiting opportunities when they present themselves".*

In the so-called "blackboard"[5] systems, such as Hearsay II (Erman et al, 1980), the processing modules called "experts" are guided by the data. In contrast to the hierarchical systems where expected solutions are made available by higher-level processes, here solutions "rise spontaneously to the surface", and are assembled into a complete interpretation as different experts complete their analysis on the basis of available data.

This type of organisation appears the most attractive of all major model types, since at least in theory, coordination becomes unnecessary. In reality however, errors and uncertainties are not checked by higher-level processes and are thus allowed to propagate across levels. To impede such error propagation, an expert module must therefore be able to question its own results at all times, and must be given the power to refine or modify its predictions. Since it cannot always make these decisions alone and at the right moment, it must be able to refer to other "experts" to resolve open issues. For example, some phonetic decisions depend on lexical and prosodic information, and vice versa, some lexical decisions depend on phonetic information. In Hearsay II, as in all "autonomous" systems, data must thus be corrected, and hypotheses must be enlarged. This is where the problem occurs. The same experts may thus be reactivated several times, which in turn can create infinite loop conditions, since knowledge sources are no longer independent of each other.

In systems with a hierarchical strategy, such as HWIN, this type of problem does not arise, but as a consequence of the hierarchical schema, the general strategy remains static — which sometimes leads to a processing inadequacy with respect to available data and to a certain lack of subtleness in the management of hypotheses and tests. Improvements thus consist of either creating a "focussing strategy", whose purpose is to activate experts that stand the greatest chance of advancing the understanding process, or of implementing an "opportunist strategy", where some form of metaknowledge is applied to the problem and to the performance of the experts.

---

[5] *In these systems, a "blackboard", or central exchange for ongoing hypotheses is maintained where experts deposit information to support or reject the hypotheses.*

The Multi-agent Models

As in general application programming, large recognition systems become close to unmanageable when constructed entirely according to traditional procedural programming principles. Object-oriented (OO) models were thus introduced in order to improve on the structure of recognition systems. From the start, the notion of "frame" (Minsky, 1975), and later, the object-oriented approach proved to be a powerful representational tool, since it offered a rich and explicit organisational structure for the various operating elements. OO techniques have since been applied extensively to so-called "high-level" problems in visual recognition and in natural language processing.

Despite the success of OO-based systems, procedural or rule-based systems have also proved to be useful in a complementary role to OO-based models (Nazif and Levine, 1984; McKeown, 1985). Procedural approaches are "action-centered", i.e. have an organisation focused on the "reasoning process" required to explain a particular activity. Such techniques have thus turned out to be successful at low-level types of processing, such as feature detection. Altogether, a variety of multi-agent architectures have been tested, representing various compromises between object-oriented and task- or rule-oriented types of processing on the one hand, and between spatial exploration, coordination and planning strategies on the other (VISIONS system, Hanson & Riseman, 1978; SIGMA system, Matsuyama and Hwang, 1985).

In terms of their theoretical orientation, these multi-agent systems are based on two different organisational paradigms, a sociopsychological and a neuropsychological paradigm. In systems based on sociopsychological notions, agents must have "beliefs" and "intentions", in order to solve a given problem in collective and coordinated fashion. These "beliefs" are cognitive operations that provide more or less well-defined representations of their environment, while the "intentions" are premises of illocutory language acts (Searle, 1969; Searle and Vanderveken, 1985) which they entertain in their communications. In this type of organisation, each agent is of sufficient size to detain a certain "intelligence" of its own. It

is therefore capable of reacting to itself and to its environment. Its behaviour is the consequence of its observations, its knowledge, its particular competence and its intentions. Currently, systems based on this type of schema have not grown much beyond ten agents because of the extensive flow of information that must be maintained between all agents of the system.

In systems based on neuropsychological notions, agents are microscopic reactive system ("feature detectors") that respond in automatic fashion to certain stimuli. They are conceived as specialised modules, similar to neural response columns or Fodor's (1983) "modules". Such systems can comprise up to several hundred agents, each of whose individual competence is rather limited, but whose global behaviour "appears intelligent" (Minsky, 1986). The advantage of this architecture is that communicative paths between such units can remain rather limited.

In Search of a "Guided Dynamic Organisation"

Both, multi-expert and multi-agent approaches to speech recognition pose some major philosophical problems. Even though it is possible to conceive of a competence distributed over a large number of agents, the emergence of "self" (in the sense of "awareness of self and its actions") is more likely the result of a centralised than a distributed organisation. This is because a structure needs a central focus of operation in order to be able to judge the value of its input processes. This appears to typify human behaviour. As we saw above, the ability to draw inferences from and to improve upon one's own performance is an important attribute of human speech recognition performance.

In fact, the weakness of a strongly distributed organisation arises from its strict separation into levels and from the entirely ascendant nature of the modules, as well as from the attendant problems of synchronous processing and inter module communication. "Detectors" in the strict sense are autonomous. To maintain this autonomy, a process must often duplicate processing that is also required elsewhere. The

alternative is waiting for another module to finish its own processing and to pass its result. However as was indicated above, this can lead to unacceptable delays or to infinite loop conditions.

So the essential difficulty is to develop a dynamic organisational structure which can profit from specialised knowledge processing encapsuled in separate layers, at the same time as it can be guided by a "stronger self" which knows when to pursue an analysis path and when to abandon it. Technically speaking, such an organisation may well be modelled by neuromimetic networks — which are probably more subtle in use and certainly more adaptive than rule-governed structures — but much more importantly, the structure must contain a considerable array of interdependent communicative paths that can be selectively inhibited by higher command. The crucial issue — one that has been insufficiently addressed by current systems and which appears to be admirably solved in human processing — is exactly how such a "higher command" and its interaction with lower modules should be structured.

Further Question Marks

Current connected-speech recognition systems are subject to a series of further critiques, some of which may again find some interesting resolutions by a consideration of the human analogy.

All current systems suffer from considerable slowness. Hearsay II, for example, was not able to meet the original goal of real-time performance, as it had been set out by the U.S. Government-sponsored DARPA project. Part of the problem again appears to be related to its centralised control (which in the case of Hearsay II operated on a single process at a time). However, even parallel processing methods have not been able to improve its speed of operation (Nii, 1986). Consider by contrast human connected-speech processing which can run well ahead of real time. Quite often, listeners can instantaneously complete well-formed sentences that are stopped in mid-stream. This indicates that their recognition processing is running minimally in parallel with the perceived

speech chain, and quite often formulates completion hypotheses well ahead of the speaker.

Some aspects of the internal functioning of current recognition systems are also "counter-intuitive", when the human analog is considered. For example, automatic recognition systems often operate by backward consultation and by the full-depth development of all alternative solutions (comparable to the typical computer chess strategy). However, there is considerable psycholinguistic evidence that for human listeners, the identification of the first word of an utterance (and even the first syllable of the first word) reduces the recognition uncertainty for the remainder of the sentence, and that human processing proceeds largely left-to-right. Furthermore, sentence complexity does not seem to be related in any systematic fashion to the type of reasoning used, which appears more in tune with the capacities of the human system which is characterised by a relatively limited short-term memory span (i.e. one type of processing where computers do better than humans). Clearly, if computers are to aspire to human levels of speech recognition performance, the implications of choosing a non-human like processing architecture must be weighed carefully.

Challenges

It is evident from this short overview that there is considerable richness in the human model. Only few of the challenges mentioned here are covered in the succeeding chapters. Nevertheless, there is ample material for some initial discussion of several of the central issues of speech synthesis and speech recognition of the coming years.

The section begins with an article by Boë, Schwartz and Vallée who address the question of how the human vowel space is subdivided in human speech processing. This problem is representative of a whole class of typical speech processing problems, since it poses the problem of dealing with the high variability of speech parameters. For example, formant values for a given vowel are not always the same, nor are they rigorously differentiated from those of another vowel, particularly in fast speech, and thus pose major detection difficulty to acoustico-phonetic input processors. The authors of

this chapter present the current thinking and some original experimentation to determine how this problem is to be addressed.

The next chapter by Gabioud takes us to the core of the question of how a direct human analog of a speech processing unit could be modelled. The relatively complex chain of processes and calculations required to produce speech synthesis by a direct model of speech articulation is described. Gabioud indicates that the current limits of this approach are in part rooted in the circumscribed availability of precise empirical information about human articulatory behaviour, and thus points out one of the characteristic weaknesses of direct models of human speech behaviour.

The chapter by Perrier and Ostry addresses the question of the direct model of articulation from a dynamic perspective. They indicate that models rooted in the study of human movement can be generalised to speech articulation. They demonstrate that without being complete algorithmic descriptions of speech articulation, such models can nevertheless offer some interesting perspectives on traditional areas of difficulty in the simulation of human speech, such as how to handle the so-called "vowel reduction" phenomena occurring in fast speech.

The chapter by Local illuminates another problem of developing synthesis devices from an interesting and unusual perspective. The question is how to model the time stream of the considerable number of parameters in a formant synthesiser (42 parameters in Klatt's original synthesiser, around 60 parameters in current, updated versions). Issuing from principles rooted in phonetic science and phonology, Local develops an original theory (whose validity is demonstrated by a fully operational implementation) of how to structure these parameters over the duration of a complete syllable.

With Geneviève Caelen-Haumont's chapter, an entirely different set of questions is raised. At issue is high-level control over prosodic parameters. Do human speakers place emphasis and intonational contours in terms of their syntactic knowledge, or do they apply semantic and pragmatic principles? Caelen-Haumont's data suggest that both principles have their importance, probably at different points in the development of a theme. Even though the transfer to

computational algorithms appears fraught with particular difficulties in the case of such higher-level control (especially in the case of semantic or pragmatic principles), the demonstration of their relevance to speech organisation poses a valid challenge for developers of natural-sounding speech synthesis devices.

The next two chapters are probably best considered as a pair, since they both address the issue of how speech can be identified against a competing background of speech or noise. Both issue from the same base literature, yet offer complementary sets of solutions. Cooke and Brown first go over the basic issues of sound separation and then enter into a detailed description of the principles that separating devices can follow to attain similar performance. Summerfield and Culling provide detailed illustrations of two processes which illustrate the advantages of exploring the human model. They arrive at logical, yet computationally surprising solutions that not only appear to do the job of separating speech sources, but also minimise the effects of potential hearing loss and exploit the partial duplication and complementary information available from the two acoustic inputs arriving via the two ears.

The chapter by Jean Caelen closes out the volume with a detailed illustration of the problems that arise when speech recognition is integrated into an operational application (in this case, a simple graphics editor) that can also respond to mouse and keyboard input. The problems surrounding this integration run the gamut from how to define the extent of a series of objects, to how to reconcile the different ways of pointing to a set of objects by speech and by mouse action. Caelen's article demonstrates that issues such as these will require a whole new set of conventions and intelligent application design for their successful resolution.

Altogether, the motivated reader will find a rich set of materials here which amply demonstrates the dynamic vibrancy of current efforts of bringing speech to the computer.

References

Erman, L.D., Hayes-Roth, F., Lesser, V.R., & Reddy, D.R. (1980). The Hearsay-II speech understanding system: Integrating knowledge to resolve uncertainty. Computer Survey, 12, 213-253.

Fodor, J.A. (1983). The modularity of Mind: An essay on faculty

psychology.Cambridge (Mass.): MIT Press.

Gong, L.Y., & Haton, J.P. (1987)A society of specialists for speech understanding IEEE-ICASSP, San Diego.

Hanson, A. R., & Riseman, E. M. (1978). VISIONS: a computer vision system for

Hayes-Roth, B., & Hayes-Roth, F. (1979). A cognitive model of planning. Cognitive Science,3, 275-310.

interpreting scenes, inComputer Vision System(pp. 303-334), New York: Academic Press.

Matsuyama, T., & Hwang, V. (1985). SIGMA: A framework for image understanding-integration of bottom-upand top-down analyses. Proceedings IJCAI, 2,908-915.

McKeown, D. M., Wilson, J. R., & McDermott, J. (1985)Rule-based interpretation of aerial imagery. IEEE Transactions of Pattern Analysis and Machine Intelligence, PAMI-7,70-585.

Minsky, M. (1975). A Framework for RepresentingKnowledge. In P.H. Winston (Ed.),The Psychology of Computer Vision(pp. 142-157). New York, N.Y.: McGraw-Hill.

Minsky, M. (1986).The society of mindCambridge, MA: MIT Press.

Nazif, A. M., & Levine, M. D. (1984). Low level imagesegmentation: An expert system. IEEE Transactions in Pattern Analysis and Machine Intelligence, PAMI-6,555-577.

Nii, H.P. (1986). Cage and poligon: Two frameworks for blackboard-based concurrentproblem-solving. Technical Report KSL-86-41 Stanford University.

Sabah, G. (1988).L'intelligence artificielle et le langageParis: Hermès.

Sabah, G. (1990). A model for interaction between cognitive processes. Proceedings of COLING'90, Vol. 3, Helsinki, pp. 446-448.

Sacerdotti, E.D. (1977)A structure for plans andbehavior. New York: Elsevier.

Searle, J. R., (1969)Speech acts.Cambridge: Cambridge University Press.

Searle, J.R., & Vanderveken, D(1985). Foundations of illocutionary logic. Cambridge University Press.