Second Edition

ENCYCLOPEDIA OF ARTIFICIAL INTELLIGENCE

Volume 2
M-Z

Awarded

American Library Association's

Outstanding Reference Source

Association of American Publishers Award

Best New Professional and Scholarly Publication

Stuart C. Shapiro
Editor-in-Chief

Speech

Edward J. Briscoe, SPEECH UNDERSTANDING, 1552-1559

From Shapiro, Stuart C., Editor-in-Chief, Encyclopedia of Artificial Intelligence, 2nd Ed., John Wiley & Sons, Inc., NY, 1992.

SPEECH UNDERSTANDING

Speech understanding is usually defined as a transduction from an initial acoustic representation of speech to a representation of meaning. For the purposes of practical systems, meaning can be defined operationally as that representation from which actions performed by the system are derived (Newell, 1975). Speech understanding is distinguished from speech recognition (qv), where the goal is to relate an utterance to (a sequence of) unique words in a dictionary. Until the early 1970s, most research focused on recognition. The five-year ARPA-funded speech project that began at that time made understanding, rather than recognition, the primary research goal. It was felt that a system's ability to respond intelligently to speech was a more meaningful criterion for the evaluation of speech systems. In addition, it was believed that the speech signal was an impoverished source of information, and knowledge of the context of an utterance was essential for its successful recognition and interpretation. Speech-recognition systems based on dynamic programming, pattern-matching techniques have been developed for utterances that consist solely of isolated words chosen from a small vocabulary, and to a lesser extent, the same techniques have been extended to connected sequences of words Rabiner and Levinson (1981). However, this approach, which works by finding the best match between variably pronounced words and a vocabulary of stored acoustic templates for words, is less suited to connected speech because the acoustic input in this case cannot be modeled effectively as a simple concatenation of the pronunciations of its constituent lexical items. In connected speech much of the variability that is factored out by pattern matching conveys information useful for both recognizing and interpreting the utterance. Therefore, it is necessary to start with more basic linguistic units than words, such as phonemes or distinctive features, and to preserve information concerning the timing and duration of the utterance. Once this step is taken, a knowledge-based, rather than pattern-matching, approach to speech processing becomes inevitable because to derive advantage from the recognition of a particular linguistic unit in the signal, it is necessary to know how that unit relates to the rest of the language in question.

Almost by definition, speech-understanding systems (SUSs) operate with connected, phrasal, sentential, or even paragraph-sized chunks of speech because "understanding" isolated words can only mean the essentially trivial process of associating some meaning with each word of the system's vocabulary and accessing this when the word is recognized. Understanding connected speech is a very complex task, and the design of SUSs has been influenced by research in fields as diverse as acoustic signal processing, (neuro)physiology, (psycho)linguistics, and psychology, as well as AI. SUSs can be classified along several dimensions; eg, number of speakers and dialects accepted or coverage of target language. To date, SUS have been built that understand a handful of speakers of similar dialect, producing a grammatically restricted subset of language with a vocabulary of about a thousand words. Although there are many potential applications for SUSs, their performance and reliability is still too poor for the majority of these to be practical. By contrast, speaker-dependent, isolated word-recognition systems for small vocabularies using whole-word pattern matching have been employed in a variety of applications, such as airline-baggage handling. Nevertheless, it is generally acknowledged that improvements to even this type of system, such as bigger vocabularies or greater speaker independence, will require a more knowledge-based approach.

THEORETICAL BACKGROUND

The transduction from speech to meaning must be mediated by a variety of components that utilize diverse knowledge sources (KSs) because the speech signal encodes, in a highly compressed and integrated fashion, many different types of information relevant to the recovery of meaning. This knowledge-based approach contrasts with that taken in whole-word template-matching systems; variability in the pronunciation of words in connected speech is no longer seen as a hindrance to pattern matching but rather as an important source of information, eg, concerning the location of word boundaries (Church, 1983) or of contextually important (stressed) information in the utterance. Figure 1 illustrates one possi-

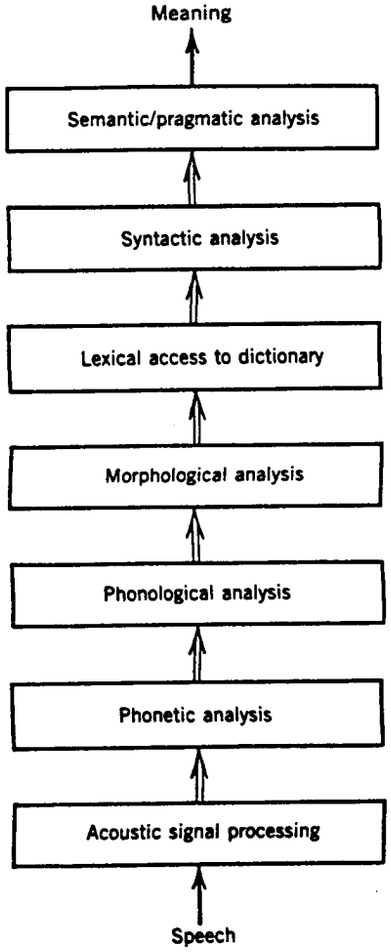

Figure 1. A typical SUS architecture.

ble organization for a SUS and the major KSs it requires to function effectively. In this organization SUS information flows upward as each component constructs intermediate representations, encoding (partial) hypotheses about the input, on the basis of the type of knowledge available to it. Acoustic signal processing digitizes the speech at a sampling rate that preserves the acoustic cues relevant to its comprehension. It also transforms the digitized signal in various ways to represent these cues in a form amenable to phonetic decoding (Flanagan, 1972; Rabiner and Schafer, 1978). For example, a spectral analysis will probably be performed and, for each analyzed frame, additional parameters, such as fundamental frequency or spectral center of gravity, computed. The parameterized signal can then be labeled as a discrete sequence of phones by searching for combinations of acoustic features. For example, if the spectral frame contains areas of "fuzziness," ie, low amplitude signals spread evenly across the spectrum, the sound is probably part of a fricative such as [f] or [v]. If the frame has a value for fundamental frequency, it must be a voiced fricative such as [v], etc. In addition, each phone is marked with suprasegmental features representing pitch, duration, and amplitude. The acoustic-phonetic transformation is crucial for the effective operation of a SUS but is still one of the least-understood aspects of speech processing. It was identified as the chief weakness in the five ARPA-funded SUSs developed in the 1970s (Klatt, 1977). Following the acoustic-phonetic transformation, a phonological analysis s performed on the phonetic representation, which identifies the linguistically important distinctions represented in the phonetic representation of the utterance, eg, levels and locations of stress, intonation contour, syllable structure, and the sequence of phonemes underlying the utterance (Lass, 1984; Ladd, 1980). Phonological analysis is essential to lexical access, which is the process of matching the phonetic form of the utterance with the canonical phonemic representations of words in a dictionary to recover the information stored there about their morphological, syntactic, and semantic properties (see MORPHOLOGY; PARSING). It undoes the effects of phonological processes such as assimilation or contraction, which apply in fluent speech; for example, the words "did" and "you" might be represented in the dictionary as the following sequence of phonemes: /dId/ and /ju:/. However, the acoustic-phonetic transformation might recover actual sounds, or phones, such as [dIǰə]; to relate this phonetic sequence to the canonical phonemic representations of "did" and "you," it is necessary to recognize that palatalization has occurred at the word boundary, changing [dj] into [ǰ], and that the unstressed vowel of "you" has been reduced to schwa. Similarly, phonological knowledge concerning allowable sequences of phonemes in syllables, often called phonotactic constraints, can be used to recognize syllable, and hence, word boundaries; for instance, in /hŏumhelp/ there must be a boundary between /m/ and the second occurrence of /h/ because no syllable in English can contain /mh/. Comparatively simple information of this type, when combined with lexical knowledge, makes the notoriously difficult task of reliably recognizing word boundaries in the acoustic signal much more tractable.

Once phonological analysis is complete, further processing of the input will be very similar to text understanding (see NATURAL LANGUAGE UNDERSTANDING). Processing will be entirely in terms of discrete symbol sets from this point upward through the SUS, and therefore, it is tempting to divide a SUS into a "recognition" phase and an "understanding" phase. However, this view is mistaken because the further morphological, syntactic, semantic, and pragmatic analysis of the utterance contributes to recognition by exploiting more of the redundancy, in the information-theoretic sense, in speech. In some of the ARPA projects, eg, there was a heavy reliance on syntactic analysis to rule out word hypotheses on the basis of syntactically inadmissable sequences. Before words hypothesized in the speech signal can be matched to lexical entries in the system's dictionary, some morphological analysis will be necessary to relate inflected variants of words to their base forms (Matthews, 1974). Apart from regular infectional morphology, such as plural /s/ or /z/, there are productive derivational morphological processes that cannot be dealt with exhaustively by expanding the number of dictionary entries; for example, there is no principled limit to the number of times "great" can be collocated with "great-grandmother" to produce a new compound noun. After morphological analysis the resulting morphophonemic representation of the speech input can be looked up in the system's dictionary to obtain syntactic and semantic information about the sequence of words hypothesized. Syntactic, semantic, and pragmatic analysis are substantially the same for speech and text understanding. However, there should be interaction between these and lower levels of analysis not only because they will contribute to correct recognition of the utterance but also because aspects of phonological analysis, particularly that relating to stress and intonation, will contribute to its interpretation. Stress, eg, is relevant to the identification of contextually new information and to finding the correct referents for pronouns. Thus, the degree of integration and interaction between different sources of information in the speech signal prevents any principled distinction between recognition and understanding. Indeed, the separation of the development of the recognition and understanding components of the SUS developed jointly by SRI International and System Development Corporation arguably explains why this major ARPA-funded SUS never worked as an integrated system (Barr and Feigenbaum, 1981).

This brief description of the contribution of different KSs to speech understanding only covers the major processes; eg, in a SUS intended to cope with several speakers, variations of voice quality, accent, and dialect must also be dealt with by the acoustic, phonetic, and phonological components. Below some issues in the design of components for the lower levels of speech analysis are discussed. The KSs deployed in speech understanding are primarily linguistic in nature, and research on them is mainly the concern of linguistic theory. However, the effectiveness of a SUS depends as much on using these KSs efficiently as on developing their content. The major contribution of AI has been to develop techniques for the representation, interpretation, and integration of KSs in

a SUS. The task of speech understanding is sufficiently complex to strain the limits of current computing technology. In existing SUSs the generally errorful nature of acoustic-phonetic analysis, and the consequent unreliability of many of the specific hypotheses under consideration by a SUS at any given point, coupled with the frequent genuine ambiguity of speech with respect to any given KS, make issues of system organization and processing strategies crucial for the construction of an effective SUS capable of functioning within practical time and space constraints. Some representative solutions that have been proposed for these problems are discussed in the following sections.

ACOUSTIC-PHONETIC ANALYSIS

Undoubtedly the most crucial area in speech processing in need of more research is acoustic-phonetic analysis. If acoustic-phonetic analysis is errorful, false hypotheses will propagate through a SUS, causing much unnecessary computation and, in the worst case, an incorrect analysis. However, if the initial phonetic representation(s) derived from the acoustic signal could be guaranteed to be unique and correct, the only indeterminacies a SUS would face would be those arising from genuine linguistic ambiguities, most of which are temporary indeterminacies resolvable in terms of further information available in the speech signal. The segmentation and identification of the acoustic signal into a sequence of linguistic units has proved extremely difficult. First, speech is a code, not a cipher (Liberman and co-workers, 1967); in other words the acoustic cues associated with segments are not in a one–one relationship with those segments; rather, these cues are heavily influenced by neighboring segments and so signal the presence of several segments in parallel. For example, the spectral cues to the presence of /d/ in /di/ and /du/ are very different because they are influenced by the following vowel. Moreover, it is not possible to divide the acoustic signal into a /d/ and a following vowel in any motivated manner. These observations prompted the theory that the invariant aspect of these segments is an abstract articulatory target that is not always achieved because of the continuous motion of the vocal tract and led to accounts of speech perception as a process mediated by speech production (Liberman and co-workers, 1967; Halle and Stevens, 1964). Such analysis-by-synthesis or completely top-down models (see PROCESSING, BOTTOM-UP AND TOP-DOWN) would be, however, very computationally expensive since they require that a SUS has the capacity to generate, in principle, all possible utterances and test them against the acoustic input. More recently, it has been argued that acoustic cues to distinctive features (Lass, 1984), as opposed to phonemes or allophones, do contain invariant cues (Stevens and Blumstein, 1978), but this claim is controversial (Lisker, 1985). Second, acoustic cues are often very minimal in unstressed speech and contexts where there is more redundancy in the speech signal. This often causes many false hypotheses in systems where the acoustic-phonetic component will hypothesize a segment from a fixed inventory, say, an allophone or allophones, for every portion of the utterance. An alternative and more attractive approach is not to force an overly specific hypothesis but to iteratively refine the analysis of the acoustic signal from broad to detailed phonetic units as far as the signal allows (Johnson and co-workers, 1985). Thus, false hypotheses will not be propagated through the SUS, although at points the phonetic analysis may lack detail. Third, the acoustic cues to units vary from speaker to speaker because of physiological differences in the vocal tract, differences of characteristic voice quality, etc (Laver, 1980). Human listeners are able to compensate for these differences rapidly and fluently, but there is still little understanding of how to model this process automatically. Most commercial speech-recognition systems require lengthy training sequences with users repeating each word in the system's vocabulary several times and are therefore very speaker-dependent. In the ARPA projects several of the SUSs developed achieved a degree of speaker independence by attempting to parameterize acoustic-phonetic analysis for a new speaker on the basis of a training sentence the system knew and the user was required to speak. A mapping could then be made between portions of the utterance and a phonetic inventory.

In addition to segmentation into some inventory of units, phonetic analysis must include a representation of the prosodic, suprasegmental aspects of speech, such as stress and intonation. The acoustic cues associated with these phenomena are fundamental frequency, duration, amplitude, and pausing. Reliably measuring fundamental frequency is difficult, as is factoring out the effects of intrinsic fundamental frequency and duration of segments from genuine suprasegmental phenomena in order to recognize stressed syllables, intonational contours, and intonational phrasing. In all of the ARPA project SUSs, suprasegmental acoustic-phonetic analysis was virtually nonexistent and segmental analysis inadequate. The final performance of each system was mainly determined by the effectiveness of higher levels of analysis in correcting errors at the phonetic level. Thus, constraints on the microworld in which each SUS operated and on the range of constructions accepted were exploited in syntactic and semantic analysis to predict what was being said. More recent systems employ more sophisticated acoustic-phonetic analysis, integrating information from a variety of transformations of the acoustic signal and constructing several types of phonetic representations, but performance is still limited to an average 70% successful recognition of phonemes from utterances produced by a small number of speakers (De Mori and co-workers, 1983).

PHONOLOGICAL ANALYSIS

Phonology is concerned with the linguistically significant, meaningful patterns of sound in a particular language, including the linguistically significant aspects of suprasegmental, prosodic phenomena. A phonological component is essential for any knowledge-based connected speech-processing system because the system will require knowledge of the phonological processes active in the language, and their domain of application, to recover canonical pronunciations for words that can be matched against a dictionary entry, and to derive further cues to the syn-

tactic and semantic/pragmatic interpretation of the utterance. Phonological components were developed for the ARPA project SUSs and other systems developed during this period (Cohen and Mercer, 1975). However, they were largely restricted to lexical, segmental processes and mostly dealt with phonologically governed variation by generating alternative pronunciations for individual lexical items and storing these in an expanded dictionary. This approach cannot deal adequately with phonological processes that span word boundaries, such as palatalization described above (Klatt, 1980). The largest domain of application for a phonological rule is the intonational phrase, which is often coextensive with a full sentence; therefore, phonology cannot be treated in terms of variant pronunciations for lexical items. Because phonological processes are rule-governed and part of the language system, a phonological analysis provides much important information for a SUS; for example, different types of phonological rule are blocked by different types of linguistic boundaries between segments, so the nonapplication of a phonological rule in an appropriate segmental environment is a clue to the presence of a boundary that blocks its application. As argued above, this is useful to aid segmentation of speech into syllables and words, but it can also provide clues for syntactic analysis; palatalization spans word boundaries but is blocked at the boundaries of major syntactic constituents (Cooper and Paccia-Cooper, 1980) so its nonoccurrence can be used to resolve an ambiguity concerning the presence of such a boundary at that point in the speech signal. Phonological rules also vary between dialects; therefore, a SUS capable of understanding speakers of different dialects would require knowledge of these differences and an ability to reconfigure itself for their speech. Palatalization, eg, occurs more frequently and more freely in dialects of American, than British, English.

At the time of the ARPA project, phonological theory was stagnant, and in particular, there was little interest in extending the domain of injury beyond segmental processes. However, since the late seventies a number of new approaches to phonology have been developed, such as autosegmental, metrical, and dependency phonology, which take as their central concern suprasegmental phenomena (Smith and van der Hulst, 1982). Few of these developments have been incorporated into SUSs, although some have been incorporated into speech-synthesis systems (Pierrehumbert, 1981; Williams, 1985) and much of this work is precise and formal enough to be suitable for machine applications. Improvements in the performance of SUSs will certainly require that these developments be incorporated into a much enhanced phonological component that can provide more than variant pronunciations for individual words.

KNOWLEDGE-SOURCE INTERPRETATION

A KS is of no use in a SUS if the knowledge it encodes cannot be represented in a way that allows its interpretation and deployment by machine. The notation employed to represent knowledge in a given field is most naturally determined by the experts in that field of knowledge; for instance phoneticians typically use the International Phonetic Alphabet for phonetic labeling. However, since choice of representation affects the application of knowledge, the representation systems of KSs in SUSs have often been a compromise between descriptive adequacy and computational efficiency. For instance, in the ARPA project every SUS, with the exception of HWIM (Woods, 1980), employed a syntactic representation thought to be unable to express all of the grammatical possibilities of English. Formal language and automata theory offer efficient algorithms for the application of KSs expressed as sets of rules with the appropriate formal properties (Aho and Ullman, 1972), and much research on representations for KSs, both in theoretical linguistics and AI, has attempted to develop descriptively adequate notations with these formal properties. For example, minimally augmented context-free notations have been argued to be descriptively adequate for English syntax (Gazdar and coworkers, 1985) and phonology (Church, 1983). Similarly, finite-state transducers have been developed for morphological analysis (Koskenniemi, 1984). However, successes of this kind do not lead automatically to computationally tractable KSs since the rule sets required to express knowledge in this form may be extremely large. In addition, it seems unlikely that all KSs employed in a SUS can be expressed within such restricted notations; therefore, more specialized and powerful techniques have also been developed, such as interpreters for production systems (McCracken, 1981) (see RULE-BASED SYSTEMS) and augmented transition networks (ATNs) (Woods, 1975) (see GRAMMAR, AUGMENTED-TRANSITION-NETWORK). Some expert-system shells (see EXPERT SYSTEMS) appear to have promising applications for the acoustic-phonetic transformation because of the more inferential and therefore principled nature of rule application (De Mori and co-workers, 1983) and the ability to factor different aspects of knowledge, which will aid parameterization of the system for different speakers (Thompson, 1984). In addition, other AI techniques associated with knowledge representation and text understanding are relevant to interpretation of KSs in a SUS. The better the understanding of a particular domain, the greater the chance of representing that knowledge both adequately and efficiently. Moreover, it is likely that different representation schemes will be most effective for different KSs; therefore, SUS architectures that impose a uniform scheme on all KSs, such as HEARSAY-II (qv) (Erman and Lesser, 1980) or HARPY (qv) (Lowerre and Reddy, 1980), are not ideal.

Choice of representation is affected by factors other than the availability of an interpretation technique for a particular scheme; for example, several SUSs do not attempt to map directly between the acoustic signal and the phonetic alphabet but construct intermediate representations, marking acoustically salient features such as nasality, to aid the process of recognizing individual phones. This reflects the difficulty, discussed above, of relating phones to distinct and invariant sets of acoustic features and a trend away from pattern matching against a continuous representation of speech toward processing of discrete symbol sets as early as possible in this process. Representations are also affected by the order in which different KSs are brought to bear on the speech signal and

the overall architecture of the SUS; recently, it has been proposed that initial phonetic analysis should mark consonants, vowels, and stressed and unstressed syllables and that this simple representation should be used to derive a set of word candidates from a suitably organized dictionary (Huttenlocher and Zue, 1984). Detailed phonetic analysis would then be applied to the stressed syllable(s) to discriminate between candidates. In a SUS employing this approach, lexical constraints are applied before detailed phonetic analysis; therefore, the role of phonetic analysis is redefined quite radically.

SYSTEM ARCHITECTURE

The bulk of the AI literature on SUSc concerns intercomponent communication during processing. This issue is crucial because ambiguities need to be resolved rapidly to avoid unnecessary computation and because redundancy between KSs can be used to factor out false hypotheses caused by either system errors or genuine ambiguity in the speed signal. For example, the acoustic-phonetic component might hypothesize an aspirated /p/ or /b/ followed by a vowel and /t/, which would result at least in the word candidates "put" and "but." However, it is likely that one of these could be rejected on the basis of syntactic analysis since verbs and conjunctions do not occur in the same syntactic environments. Similarly, there might be a genuine syntactic ambiguity in an utterance, such as "He gave her dog biscuits." in which "her" may be functioning as an adjective or noun. But in this case the ambiguity can be resolved by the different stress and intonation that will accompany the two interpretations. The architectures proposed are basically hierarchical, like that of Figure 1, with a serial flow of information through a chain of component KSs, or heterarchical with no constraint, in principle, on the flow of information between components (Reddy and Erman, 1975). The advantage of the hierarchical approach is that there appears to be a natural order for the application of KSs to speech input; syntactic analysis can only proceed on the basis of lexical information, etc. Moreover, overall system control is simple. However, there are many occasions when nonserial interactions between the chain of components are useful; for example, aspects of the prosodic, suprasegmental structure of an utterance will be relevant to its phonological, syntactic, semantic, and pragmatic interpretation. Nonserial interaction can be achieved within the hierarchical model by passing up all of the possible analyses compatible with a given component to the next component, which then selects a subset of analyses, etc. But this only works if the intermediate representations passed up through the SUS are enriched to include all of the information analyzed so far that may be of use to some higher component; thus, the input to the syntactic component, in addition to syntactic information about words, must include all of the available information of potential relevance to syntactic analysis, such as prosodic information, and all information relevant to semantic/pragmatic analysis must be carried through as well. This is likely to strain representation schemes and is computationally expensive because many unnecessary, false hypotheses are computed. These false hypotheses are often avoidable, in principle, because the disambiguating information is temporally available, encoded in the part of the speech signal already analyzed by lower levels, but in the hierarchical model it is not applied until this input reaches the appropriate component in the serial chain. Heterarchical systems avoid this inefficiency by allowing components to apply in the most efficient order for a given input at the expense of a very complex flow of control within the system and considerable intercomponent communication complexities. Each component must be provided with the means to request and receive information from, or start specific processing in, any other component. This requires specialized communication channels between every component in the system. Developing an adequate control system for such a model may well be impossible because it involves envisaging all possible flows of control at the design state. Attempts at developing such models have been reduced to human simulation of each component (Woods and Makhoul, 1973). In practice, workable heterarchical models for SUSs have been restricted to uniform representations across KSs and a single global data structure, as in blackboard systems (qv).

Since the ARPA project there has been much interest in SUSs that can be run on parallel machines (Fahlman and co-workers, 1983) and the emphasis has been on hierarchical systems that still achieve interaction through selective filtering of hypotheses. Thus, a simple architecture is maintained and more powerful hardware used to cope with the extra computation. One recent major project proposes to use the chart (Kay, 1973) as a global data structure within a hierarchical architecture, employing as much parallelism as the linearized nature of speech input will allow (Thompson, 1984). This approach shares the advantage of a global data structure with the blackboard architecture but does not require a uniform representation scheme across components. None of these proposals represents a theory of nonserial interaction in speech understanding; rather, they offer general architectures that attempt to support any interaction that may be required. The designers of the blackboard specifically wanted an architecture capable of supporting arbitrary interactions and thus application to other tasks, such as vision (Erman and Lesser, 1975). An alternative approach is to specify explicitly the interactions required in a SUS and to develop a specialized architecture capable of supporting them; this approach requires a better understanding of speech than is current but offers the possibility of far more efficient and effective SUS architectures (Briscoe and Boguraev, 1984).

PROCESSING STRATEGIES

Various processing strategies have been imposed on different SUS architectures in an attempt to reduce the computation required for successful analysis in the normal case. Both hierarchical and heterarchical systems can operate bottom-up, in an essentially data-driven way, or top-down, using knowledge to produce hypotheses concerning

the input (see PROCESSING, BOTTOM-UP AND TOP-DOWN). However, most recent SUSs have operated bottom-up because of the rather weak predictability of speech on the basis of the KSs that can currently be deployed effectively in a SUS. Similarly, SUSs can explore the search space in a depth-first or breadth-first manner (see SEARCH; SEARCH, DEPTH-FIRST). Most have operated breadth-first because of the uncertain and errorful nature of many hypotheses but have employed scoring techniques to keep the size of the active search space manageable. One such technique, shortfall scoring, which involves measuring the summation of individual word candidate scores against a theoretical upper bound for this score and process the hypothesis with the least difference first, guarantees that a SUS will find the best scoring complete hypothesis for the utterance first (Woods, 1982). However, this does not guarantee that the highest ranked hypothesis is correct; the effectiveness of the components that contribute to the generation of word hypotheses is still the crucial factor in the overall performance of the system. The scoring of partial hypotheses throughout a SUS in a more linguistically motivated fashion is extremely difficult. Scores must be carried across components and should reflect the differing contributions of each KS. However, the weight that should be attached to any KS must vary with context; for instance, in the recognition of an unstressed and phonetically reduced preposition, syntactic analysis should be weighted more highly relative to acoustic analysis than in recognition of a stressed syllable. In addition, analyses must be scored through time; an analysis that starts with low scores may end with the highest because redundancy in the information encoded in speech propagates both left and right through the input. Although some scoring schemes that have been used in implemented SUSs do improve performance, this is either for theoretical reasons connected with the scoring technique, as with shortfall scoring, or because they have been developed by trial and error and evaluated solely on the basis of run-time performance, as with the focus-of-attention mechanism in the HEARSAY-II blackboard system. In the former case the technique is useful but limited; in the latter case potentially valuable insights into the task of speech understanding become lost in the scoring technique (Hayes-Roth, 1983).

Analysis of the speech signal can proceed from left to right through the linearized signal or middle out in both directions from islands of greater acoustic reliability. This island-driven approach has the advantage of taking relatively error-free phonetic data as its starting point at the expense of a more complex control structure and system organization, as in HWIM (qv) (Woods, 1982). Human listeners appear to pay greater attention to stressed syllables (Cutler and Norris, 1985), which are generally more clearly enunciated and therefore more easily analyzed phonetically. In addition, the phonological structure of English vocabulary is constrained in such a way that a unique word can usually be derived from a crude phonetic analysis of syllable structure coupled with detailed analysis of its stressed syllable (Huttenlocher and Zue, 1984). Therefore, the island-driven approach is essentially correct, although it would be more effective if processing began at stressed syllables by explicitly searching for them rather than at arbitrary high scoring portions of the acoustic signal. A related approach that avoids the extra overheads in HWIM caused by middle-out analysis is to use an appropriate scoring strategy in a left-to-right system, which is able to take account of forthcoming speech events in the right context of the existing partial hypothesis (Johnstone and Altmann, 1984).

CURRENT TRENDS

Since the ARPA project in the seventies there has been a period of problem-oriented, rather than system-building, research in speech understanding. Much of this research has focused on the acoustic-phonetic transformation as a result of new evidence demonstrating the informational richness of the acoustic signal (Cole and co-workers, 1980). Now there is renewed interest in building complete systems (Thompson, 1984) incorporating this research and renewed concern with issues such as system architecture. However, the majority of the knowledge-based systems that are being developed are restricted to continuous speech recognition rather than understanding. Improvements in acoustic-phonetic analysis suggest that higher levels of analysis are not crucial for recognition of continuous speech, contrary to prevailing opinion at the time of the ARPA project. In addition, the problems of understanding, such as knowledge representation (qv) issues, the restriction to a microworld, etc, remain unsolved.

SYSTEMS

The main SUSs developed in the ARPA project were HARPY, HWIM, HEARSAY-II, and SRI/SDC. HARPY came closest to the performance criteria specified for the project (Newell and co-workers, 1973). However, HARPY's architecture required precompilation of all KSs into a single finite-state network so the language accepted by the system was more restricted than that for the other systems (Hayes-Roth, 1983). Each of these systems is briefly described by its chief designer in Lea (1980) and evaluated in Klatt (1977) and Barr and Feigenbaum (1981). A system of the same period developed at IBM but with improved performance over HARPY on the same subset of English is described in Bahl and co-workers (1976). The HEARSAY-II system is reimplemented as a production system in McCracken (1981), and extensions and improvements to the original blackboard architecture are described in Hayes-Roth (1983). Several SUSs have been developed for European languages, such as KEAL (Mercier and co-workers, 1980) and MYRTILLE-II (Pierrel and Haton, 1980) for French and EVAR (Niemann, 1982) for German. However, these systems have not surpassed the ARPA project systems in performance or design. An automated airline reservation system that incorporates continuous speech understanding is described in Levinson and Shipley (1980). This system, developed at Bell Laboratories, conducts a dialogue over a telephone to establish the appropriate reservation. It employs whole-word template-matching techniques to recognize words

from a 127-word vocabulary but relies on semantic constraints deriving from this very restricted task domain and syntactic constraints imposed by a restrictive finite-state grammar to achieve robust performance.

FURTHER READING

The best introductions to speech understanding are Lea (1980) and Reddy (1975). Two further more recent collections are Cole (1980) and Simon (1980), and Fallside and Woods (1985) provide an up-to-date, advanced course. Articles on the higher levels of speech understanding can be found in the journals *Artificial Intelligence* and *Computational Linguistics*. Issues relevant to acoustic-phonetic analysis are dealt with in *Journal of the Acoustical Society of America, Journal of Phonetics* and *Language and Speech*. Many articles on whole-word template-matching techniques for isolated and connected speech and on speech signal processing can be found in the journal of the Institute of Electrical and Electronic Engineering, Acoustics, Speech and Signal Processing. Articles on the application of speech-processing systems can be found in *Speech Technology*, and *Human Factors* sometimes contains performance evaluations of systems. The *International Journal of Man–Machine Studies* also publishes articles on speech processing. Many conference proceedings also contain relevant articles, such as the *International Joint Conference on Artificial Intelligence, COLING, Association of Computational Linguistics, Acoustic Society of America, International Conference on Acoustics, Speech and Signal Processing*, and others.

BIBLIOGRAPHY

A. V. Aho and J. Ullman, *The Theory of Parsing, Translating and Compiling*, Vol. I, Prentice-Hall, Englewood Cliffs, N.J., 1972.

L. R. Bahl, J. K. Baker, P. S. Cohen, N. R. Dixon, F. Jelinek, R. L. Mercer, and H. F. Silverman, "Preliminary Results on the Performance of System for the Automatic Recognition of Continuous Speech," *International Conference on Acoustics, Speech and Signal Processing*, Philadelphia, Pa., 1976, pp. 512–514.

A. Barr and E. A. Feigenbaum, *The Handbook of Artificial Intelligence*, Vol. I, Kaufmann, Los Altos, Calif., 1981.

E. J. Briscoe and B. K. Boguraev, "Control Structures and Theories of Interaction in Speech Understanding Systems," *Proceedings of COLING84*, Stanford, Calif., 1984, pp. 259–266.

K. W. Church, *Phrase Structure Parsing: A Method for Taking Advantage of Allophonic Constraints*, Indiana University Linguistics Club, Bloomington, Ind., 1983.

P. S. Cohen and R. L. Mercer, "The Phonological Component of an Automatic Speech Recognition System," in D. R. Reddy, ed., 1975, pp. 275–320.

R. A. Cole, ed., *Perception and Production of Fluent Speech*, Erlbaum, Hillsdale, N.J., 1980.

R. A. Cole, A. I. Rudnicky, V. W. Zue, and D. R. Reddy, "Speech as Patterns on Paper," in R. A. Cole, ed., 1980, pp. 3–50.

W. E. Cooper and J. Paccia-Copper, *Syntax and Speech*, Harvard University Press, Cambridge, Mass., 1980.

A. Cutler and D. Norris, "Syllable Boundaries and Stress in Speech Segmentation," *Proceedings of the 109th Meeting of Acoustic Society of America*, Austin, Tex., 1985, pp. S39.

R. De Mori, P. Laface, G. Petrone, and M. Segnan, "Access to a Large Lexicon Using Phonetic Features," *Proceedings of EUSIPCO, Erlangen, 1983*.

L. D. Erman and V. R. Lesser, "A Multi-Level Organisation for Problem Solving Using Many, Diverse, Cooperating Sources of Knowledge," *Proceedings of the Fourth International Joint Conference of AI*, Tibilisi, Georgia, Morgan-Kaufmann, San Mateo, Calif., 1975, pp. 483–490.

L. D. Erman and V. R. Lesser, "The Hearsay-II Speech Understanding System: A Tutorial," in W. A. Lea, ed., 1980, pp. 361–381.

S. E. Fahlmann, G. E. Hinton, and T. J. Sejnowski, "Massively Parallel Architectures for AI: NETL, Thistle and Boltzmann Machines," *Proceedings of the Third National Conference for Artificial Intelligence*, Washington, D.C., AAAI, Menlo Park, Calif., 1983, pp. 109–113.

F. Fallside and W. A. Woods, eds., *Computer Speech Processing*, Prentice-Hall, Englewood Cliffs, N.J., 1985.

J. L. Flanagan, *Speech Analysis, Synthesis and Perception*, 2nd ed., Springer-Verlag, New York, 1972.

G. J. M. Gazdar, G. K. Pullum, I. A. Sag, and E. Klein, *Generalized Phrase Structure Grammar*, Blackwell, Oxford, 1985.

M. Halle and K. N. Stevens, "Speech Recognition: A Model and a Program for Research," in J. A. Fodor and J. J. Katz, eds., *The Structure of Language*, Prentice-Hall, Englewood Cliffs, N.J., 1964, pp. 604–612.

B. Hayes-Roth, *A Blackboard Model of Control*, Report No. HPP-83-38, Department of Computer Science, Stanford University, Calif., 1983.

D. P. Huttenlocher and V. W. Zue, "A Model of Lexical Access from Partial Phonetic Information," *Proceedings of the IEEE Conference on Acoustics, Speech and Signal Processing*, San Diego, Calif., Mar. 1984, pp. 2641–2644.

S. R. Johnson, J. H. Connolly, and E. A. Edmonds, "Spectrogram Analysis: A Knowledge-Based Approach to Automatic Speech Recognition," in M. A. Bramer, ed., *Research and Development in Expert Systems*, Cambridge University Press, Cambridge, UK, 1985.

A. M. Johnstone and G. Altmann, *Automated Speech Recognition: A Framework for Research*, Research Report No. 233, Edinburgh University, Department of AI, Edinburgh, 1984.

M. Kay, "The MIND System," in R. Rustin, ed., *Natural Language Processing*, Algorithmics, New York, 1973, pp. 155–188.

D. H. Klatt, "Review of the ARPA Speech Understanding Project," *J. Acoust. Soc. Am.* **62**, 1345–1366 (1977).

D. H. Klatt, "Speech Perception: A Model of Acoustic-Phonetic Analysis and Lexical Access," in R. A. Cole, ed., 1980, pp. 243–288.

K. Koskenniemi, "A General Computational Model for Word-Form Recognition and Production," *Proceedings of COLING84*, Stanford, Calif., July 1984, pp. 178–181.

D. R. Ladd, *The Structure of Intonational Meaning*, Indiana University Press, Bloomington, Ind., 1980.

R. Lass, *Phonology*, Cambridge University Press, Cambridge, UK, 1984.

J. Laver, *The Phonetic Description of Voice Quality*, Cambridge University Press, Cambridge, UK, 1980.

W. A. Lea, ed., *Trends in Speech Recognition*, Prentice-Hall, Englewood Cliffs, N.J., 1980.

S. E. Levinson and K. L. Shipley, "A Conversational-Mode Airline Information and Reservation System Using Speech Input and Output," *Bell Sys. Tech. J.* **59,** 119–137 (1980).

A. M. Liberman, F. S. Cooper, D. P. Shankweiler, and M. Studdert-Kennedy, "Perception of the Speech Code," *Psychol. Rev.* **74,** 431–461 (1967).

L. Lisker, "The Pursuit of Invariance in Speech Signals," *J. Acoust. Soc. Am.* **77,** 1199–1202 (1985).

B. T. Lowerre and D. R. Reddy, "The Harpy Speech Understanding System," in W. A. Lea, ed., 1980, pp. 340–360.

D. L. McCracken, *A Production System Version of the Hearsay-II Speech Understanding System,* UMI Research, Ann Arbor, Mich., 1981.

P. H. Matthews, *Morphology,* Cambridge University Press, Cambridge, UK, 1974.

G. Mercier, A. Nouhen, P. Qiunton, and J. Siroux, "The KEAL Speech Understanding System," in J. C. Simon, ed., 1980, pp. 525–545.

A. Newell, "A Tutorial on Speech Understanding Systems," in D. R. Reddy, ed., 1975, pp. 3–54.

A. J. Newell, J. Barnett, J. W. Forgie, C. Green, D. H. Klatt, J. C. R. Licklider, J. Munson, D. R. Reddy, and W. A. Woods, *Speech Understanding Systems: Final Report of a Study Group,* North-Holland, Amsterdam, 1973.

H. Niemann, "The Erlangen System for Recognition and Understanding of Continuous Speech," in J. Nehmer, ed., Springer-Verlag, Berlin, 1982, pp. 330–348.

J. B. Pierrehumbert, "Synthesising Intonation," *J. Acoust. Soc. Am.* **70,** 985–995 (1981).

J. M. Pierrel and J. P. Haton, "The MYRTILLE-II Speech Understanding System," in J. C. Simon, ed., 1980, pp. 553–570.

L. R. Rabiner and S. E. Levinson, "Isolated and Connected Word Recognition—Theory and Selected Applications," *IEEE Trans. Commun.* **Com-29**(5), 621–659 (1981).

L. R. Rabiner and R. W. Schafer, *Digital Processing of Speech Signals,* Prentice-Hall, Englewood Cliffs, N.J., 1978.

D. R. Reddy, ed., *Speech Recognition,* Academic Press, Inc., New York, 1975.

D. R. Reddy and L. D. Erman, "Tutorial on System Organisation for Speech Understanding," in D. R. Reddy, ed., 1975, pp. 457–480.

J. C. Simon, ed., *Spoken Language Generation and Understanding,* Reidel, Dordrecht, Holland, 1980.

N. Smith and H. van der Hulst, eds., *The Structure of Phonological Representations,* Foris, Dordrecht, Holland, 1982.

K. N. Stevens and S. E. Blumstein, "Invariant Cues for Place of Articulation in Stop Consonants," *J. Acoust. Soc. Am.* **64,** 1358–1368 (1978).

H. Thompson, "Speech Transcription: An Incremental Interactive Approach," *Proceedings of the European Conference on AI,* Pisa, Italy, Sept. 1984, Elsevier Science, New York, pp. 697–704.

B. J. Williams, "A Metrical Algorithm for Lexical Stress Assignment in English," *Proceedings of 109th Meeting of Acoustic Society of America,* Austin, Tex., 1985, p. S39.

W. A. Woods, "Syntax, Semantics and Speech," in D. R. Reddy, ed., 1975, pp. 345–400.

W. A. Woods, "Control of Syntax and Semantics in Continuous Speech Understanding," in J. C. Simon, ed., 1980, pp. 337–364.

W. A. Woods, "Optimal Search Strategies for Speech Understanding Control," *Artif. Intell.* **18,** 295–326 (1982).

W. A. Woods and J. Makhoul, "Mechanical Inference Problems in Continuous Speech Understanding," *Proceedings of the Third International Joint Conference Artificial Intelligence,* Stanford, Calif., Morgan-Kaufmann, San Mateo, Calif., 1973, pp. 200–207.

E. J. Briscoe
University of Cambridge

SPHINX

SPHINX is the first speech recognition system that achieved high performance on a large vocabulary, speaker independent, continuous speech task. The system uses stochastic hidden Markov models (HMMs) to learn and generalize from a large multi-speaker database. SPHINX also uses techniques such as multiple codebooks, generalized triphones, and function word models. On a 1000-word task, SPHINX recognizes continuous speech from any speaker with an accuracy of 96% (see K. F. Lee, *Automatic Speech Recognition: The Development of the SPHINX System,* Kluwer Academic Publishers, Boston, Mass., 1989).

Kai-Fu Lee
Apple Computer, Inc.

SSS. . .

SSS. . . (Summarization Summarization Summarization. . .) was developed to explore the utility of the encoding relationship between semantic memory and text (see R. Alterman and L. Bookman, "Reasoning About a Semantic Memory Encoding of the Connectivity of Events," *Cogn. Sci.,* in press.) An important idea developed in the work on SSS. . . was the notion of conceptual roots. Roughly, the conceptual roots correspond to the basic notions of the narrative text: the framework in the terms of which the narrative was developed. SSS. . . determined the conceptual roots from a directed acyclic graph structure that represents the coherence of the event concepts in the text as derived from the encoding provided by semantic memory. An interesting property of the conceptual roots are that they are the minimal set that covers the semantic memory-based graph encoding of the case. SSS. . . used the conceptual roots to succinctly explain the connection between any two concept coherent events in the narrative. Also implemented in SSS. . . was a measure of importance that quantifies the conceptual emphasis of the narrative. Given this measure of importance, it was shown that each of the important nodes is either a conceptual root or covered by one of the conceptual roots. Lastly, SSS. . . combined evidence from semantic memory-based coherence graph of the case, the conceptual root analysis and the importance measure, to generate a description of the basic event content of the narrative (ie, a basic summary) [see R. Alterman and L. Bookman, "Some Computation Experiments in Summarization, *Discourse Processes* **13,** 143–174 (1990)].

Richard Alterman
Brandeis University

ENCYCLOPEDIA OF ARTIFICIAL INTELLIGENCE

This extensively revised and expanded *Second Edition* of the *Encyclopedia of Artificial Intelligence* defines the discipline by bringing together the core of knowledge from all fields encompassed by AI. It covers the latest developments in current AI topics such as neural networks, fuzzy logic, machine vision, natural language generation, and many more. Includes:

- Over 450 articles—all entries written expressly for the *Encyclopedia*
- Over 5,000 literature references; 454 illustrations and color photographs
- Over 50% new and revised material
- Exemplary indexing and cross-referencing for easy, complete information access to all topics

Praise for the *First Edition*...

"The *Encyclopedia* is a wonder of clarity and scope: surprisingly easy to read...the clarity is an especially pleasant surprise, considering the articles were all written by AI experts...It's a treasure house of easily accessible knowledge."
—*Language Technology*

"Excellent bibliographies are attached to most of the articles, and diagrams and sketches are clear and helpful. The indexing and cross-indexing are exemplary. As the editor points out, the reader will be led by the extensive cross-references to almost every other article..." —*Artificial Intelligence Reporter*

"The *Encyclopedia* is a first-class piece of work that will be an indispensable part of any AI library." —*Computing Reviews*

"...A tour de force...a truly fantastic encyclopedia which no one in the field of artificial intelligence can afford to be without."
—*Systems Research & Information*

WILEY-INTERSCIENCE
John Wiley & Sons, Inc.
Professional, Reference and Trade Group
605 Third Avenue, New York, N.Y. 10158-0012
New York • Chichester • Brisbane • Toronto • Singapore
ISBN 0 471-50307-X (Two-Volume Set)
ISBN 0 471-50305-3 (Vol. 1)
ISBN 0 471-50306-1 (Vol. 2)

ISBN 0-471-50306-1

9 780471 503064 90000