CONTRIBUTIONS OF SPEECH SCIENCE TO THE TECHNOLOGY OF MAN-MACHINE VOICE INTERACTIONS

WAYNE A. LEA

SPEECH COMMUNICATIONS RESEARCH LABORATORY, INC.
SANTA BARBARA, CALIFORNIA

ABSTRACT

Previous interdisciplinary research at Speech Communications Research Laboratory has dealt with a variety of topics in linguistics, speech physiology, perception, and acoustics, plus the interactions among those disciplines. Linear prediction and prosodic correlates of linguistic structures are two examples of research topics that have led to many practical contributions in such application areas as speech recongition. Work in speech recognition has included techniques for vowel identification and normalization, locating syllables, detecting stresses and phrase boundaries, accurately transcribing speech, developing and applying phonological rules, and participating in various aspects of the ARPA SUR project.

Currently a review of the ARPA SUR project and a survey of the speech understanding field are being conducted, with recommendations forthcoming regarding future needs. Several presentations and publications, including a forthcoming book, will report such work. Future plans include prosodics research, phonological rules for speech understanding systems, and continued interdisciplinary phonetics research. One outstanding conclusion from the current review and survey is a renewed call for improved acoustic phonetic analysis capabilities in speech recognizers.

Submitted for publication in the Proceedings of the Workshop on Voice Technology for Interactive Real-Time Command and Control Systems Application, NASA, Ames Research Center, Moffett Field, California.

1. Introduction

Speech Communications Research Laboratory (SCRL) is a non-profit research laboratory that was established on the premise that the experimental and theoretical study of spoken language is not simply an adjunct to some other discipline such as electrical engineering or linguistics, but rather it is a distinct and major field of investigation.

It is difficult and, we believe, undesirable to separate our work in speech recognition from the many other disciplines and speech communication problems with which SCRL has worked. This paper consequently begins with a review of the wide range of speech communications projects SCRL has undertaken (section 2). Rather than simply list the many projects, I have organized them within a framework which graphically illustrates the interactions between speech acoustics, physiology, and linguistics. I also offer two examples, concerned with linear predictive analysis and prosodic correlates of linguistic structures, that illustrate how techniques that are directly applicable to speech understanding systems actually originate from interdisciplinary experimental and theoretical research, and then can be turned around to offer evidence for significant changes and new efforts in theory and experimentation.

In section 3, I complete the review of previous SCRL work by briefly describing specific studies in speech recognition that have been conducted at SCRL. These include a number of modest efforts in technology development, and a large project of participation in the Speech Understanding Research ("ARPA SUR") Project sponsored in 1971-1976 by the Advanced Research Projects Agency of the Department of Defense.

Turning from past (Pre-FY '78) work to present and future (Post-FY ' 77) efforts, in section 4 I describe a current contract Dr. June E. Shoup and I are directing, to review the entire ARPA SUR project, to survey all the current technology in speech understanding, and to offer recommendations for further work. This Tri-Services sponsored contract is directly in line with the purposes of this workshop, and should be of widespread interest. We are planning to publish several papers, present several conference talks, and edit two books about speech recognition work throughout the world, and so these outcomes from our project are described in section 5. It is also our hope that from this workshop, from our review, and from related cooperative efforts can come a cataloging of available speech recognition tools, speech databases, and general laboratory facilities for speech analysis, transcription of speech, and collecting statistics about speech regularities. This I discuss briefly in section 6.

Finally, in section 7, I outline our plans for future work on speech understanding.

2. The Practical Utility Of Interdisciplinary Research

An understanding of the mechanisms and structures which underly speech is essential to effective man-machine voice communication. We need to call upon the expertise of linguists, phoneticians, engineers.

psychologists, physiologists, speech clinicians, computer scientists, and many other disciplines. For example, it was the psychologists that in recent years clearly demonstrated that no single modality of human communication is as effective in practical problem solving as speech, and that speech is the essential ingredient of the most effective multi-modality communication links (Chapanis, 1975).

Engineers and mathematicians gave us the array of valuable speech analysis tools ranging from microphones and electronic filters to Fourier analysis capabilities, fast Fourier transforms, linear predictive analysis, and many other practical devices and algorithms. Computer scientists have given us that fast and versatile tool, the general purpose digital computer, and all its special purpose versions and peripheral devices. More recently, the computer scientists and artificial intelligence advocates have given us practical and effective methods for answering the twenty-year-old call for use of higher-level linguistics knowledge (phonological rules, lexicons, syntax, semantics, and pragmatics) in speech recognition (Denes, 1957; Lindgren, 1965). Decades of work and ideas in acoustic-phonetics, articulatory phonetics, and perception have brought us the phones, phonemes, manner-and-place-of-articulation features, coarticulation constraints, and guidelines about which acoustic changes are truly important (i.e. perceptible), upon which almost all speech recognition and synthesis work is based. Prosodics, as the study of stress, intonation, and the rhythm and timing of speech, had for decades been the concern of comparatively few isolated speech scientists and language teachers, but has recently become one of the prominent subjects in work on speech synthesis and recognition. And so the listing could continue, showing repeated ways in which today's technology builds on yesterday's interdisciplinary science and creative thought. Recently, the ARPA SUR project showed that such a variety of disciplines could work together effectively to develop powerful systems that can successfully understand spoken sentences.

SCRL has, since its founding in 1966, been concerned with the scientific study of the basic linguistic structures of spoken languages, and with the application of this information to problems in electronic communication and speech automation. Gordon Peterson, Founding President and first Director of SCRL, said at the time of SCRL's formation:

> " It is the purpose of the Laboratory to provide a place where scientists and scholars from various disciplines, both technical and humanistic, can work together in mutal respect and enthusiasm on the endless and fascinating problems of speech communication. "

Since that challenging call in 1966, SCRL has been living up to its general goals of discovering basic processes underlying speech communication and sharing the resulting information in the public interest.

While it is recognized that many contributions from basic research do not have widespread impact for many years after the laboratory research is accomplished, it is SCRL's policy to do basic research with specific applications in mind. The result has been that some outstanding ideas and developments at SCRL have had almost immediate direct benefits in practical applications. Perhaps one of the best known examples would be the leading theoretical work of John Markel and his colleagues (Markel, 1972; Markel and Gray, 1973; 1974; 1976) on linear predictive analysis, which has already been applied in systems for speech recognition, speaker authentication or identification, and early detection of laryngeal cancer. Markel is currently applying his techniques to government applications in speaker recognition, within a newly formed applications-oriented company he directs. His linear predictive coding techniques have also been adopted by many other groups working on speech analysis and synthesis throughout the world. If someone had stopped that type of rigorous mathematical work at its early stages only a few years ago, on the mistaken notion that it was irrelevant to immediate practical needs, where would our speech analysis and synthesis capabilities be today? We might still be struggling to extract the really important spectral cues (formants, fundamental frequency, glottal waveforms, vocal tract area functions, et.) from the complicated, noisy speech spectra that for twenty or more years had defied reliable automatic analysis.

Linear predictive analysis is a good model for illustrating the interdisciplinary origins and applicabilities of speech research. The mathematical models, that are now implementable in practical forms in general purpose (or specialized) computers, have been shown to be appropriate to capture the essence of the accustic modulation of a vocal-cord source that is produced by the variable-cross-section vocal tract. Linear prediction permits detection of vocal tract resonances (formants or transfer-function poles), voice fundamental frequency and waveforms of airflow at the vocal cords, and radiation impedance at the lips. It is known to be appropriate for vowels and oral consonants, and even though our knowledge of articulation and acoustic phonetics suggests its mathematical inapplicability for nasal consonants, practical approximations and perceptual significances tell us that it is possible to learn something about the speech (e.g., approximate nasal resonances and bandwidths) even when the model's mathematical assumptions are not strictly met. Here we see acoustics, articulatory phonetics, perception, linguistic category distinctions, mathematics, computer science, and practical engineering approximations all coming into play. Then we see linear predictive anaylsis used to aid vowel and consonant identification in speech recognition, plus detect talker-specific differences in vocal tracts and voice sources, and even detect laryngeal cancer and other speech pathologies. One recent project at SCRL used the residual energy function

from a linear predictive analysis to detect laryngeal (voice) pathologies such as cancer, and to provide "voice profiles" that may be useful in clinical, musical, and legal applications (Davis, 1976)

Another example of interdisciplinary interactions is my own growing interest in prosodic structure. When, in 1966, Gordon Peterson and his colleagues at SCRL first introduced me to the obscure area of phonetics and linguistic studies they called "prosodic structures", I had no idea how prosodics studies would lead to such a variety of scientific questions and practical applications. Following the linguists' arguments that stress patterns are determined by the phrase structures of sentences, and the phoneticians' studies of acousitc prosodic correlates of stress, I hypothesized that one should be able to determine aspects of syntactic structure directly from acoustic prosodic features. This led to the development of a computer program which detected about 90% of major phrase boundaries in connected speech, using only fall-rise valleys in intonation patterns. Another program detected syllabic nuclei from bandlimited energy functions, and used energy, syllabic durations, and fundamental frequency contours to successfully locate about 90% of the stressed syllables. Extensive series of experiments were conducted on the intonation, perceived stress patterns, rhythms, and pauses in various speech texts. Methods were devised for using such prosodic information to aid phonemic analysis, word matching, and parsing in speech understanding systems. In fact, a general prosodically-guided speech understanding system strategy was outlined, and aspects of it were incorporated into the developing system at Sperry Univac (Lea, Medress, and Skinner, 1975).

All this prosodics research which I did while at Sperry Univac is summarized in a recent report (Lea, 1977). It clearly showed the potential for extracting aspects of syntactic structure from acoustic prosodic data, independent of any knowledge of the wording of the sentence. Prosodics also can be used to reduce the set of alternative words that should be hypothesized at each point in an unknown utterance. Hypothesized words should have stresses expected where they are actually found in the acoustic prosodic data (for example, word-finally stressed "abridge" should not be hypothesized or should be given a lower priority for testing where the prosodics clearly suggest an initially-stressed word like "average"). Also, only certain words can be in phrase-initial or phrase-final positions, so if a phrase boundary is reliable detected, one can confine hypothesized words to those that could appear in those patterns.

Those prosodic studies, which began from general linguistic theories and acoustic phonetic experiments, thus developed into substantial contributions to practical aspects of computer understanding of spoken sentences. Then, as if to complete the circle, some of the acoustic prosodic features detected in such analyses led to widespread theoretical implications, such as explanations for how tones develop or disappear in the historical change of a language (or family of languages), how

consonants interact with tones in tone languages, why stresses tend to be equally spaced (isochronous) in English, which of the linguist's stress rules are evident in acoustic data and listeners perceptions of stress, etc. I also used available automatic phonetic analysis routines to confirm a long-held notion that stressed syllables provide "islands of phonetic reliability" in speech. These studies also raised questions about the physiological origins of higher fundamental frequencies in high (vs. low) vowels, relationships between larynx height and fundamental frequency, the physiological origin of gradually falling intonation, etc.

We thus have two quite different examples of practical benefits coming from some interdisciplinary research. A detailed discussion of other SCRL interdisciplinary work is impossible here, but we can list many of the other topics that have been studied, and indicate some structure for relating all these studies to each other and to our main topic of speech recognition.

Gordon Peterson characterized the interrelationships between acoustics, physiology, and linguistics by the basic triangle shown in Figure 1. I have illustrated on the diagram the various topics of research to which SCRL has contributed during its various government-sponsored and privately funded contracts and grants. This listing of topics was compiled from the list of over 100 journal articles, book chapters, and reports, plus 14 books and monographs, that SCRL researchers have published. The work ranges from abstract linguistic studies like grammar, phonology, dialects, and abstract prosodic ("prosodemic") structures, to extensive studies of acoustic features of vowels and consonants, and a variety of signal processing techniques and applications. Physiology, as something of a "way station" between linguistics and acoustics, has been the subject of several medical studies and some mathematical modelling at SCRL.

Outstanding among the published works from SCRL are Peterson and Shoup's "Physiological and Acoustic Theories of Phonetics" (1966). These links between linguistics and either acoustics or physiology are shown by the top and left arrows in Figure 1. Also linking linguistics and acoustics are developments of dictionaries specifying the actual ways words are pronounced in various forms of communication (read speech, formal talks, conversation, etc.). Speech synthesis is an "encoding" effort, which allows going from specified linguistic messages to automatically composed acoustic forms that are acceptable and intelligible to listeners. Speech recognition, the primary topic of the remainder of this paper, is the opposite process of automatically determining linguistic messages from acoustic data.

Many researchers have noted the difficulty of relating accoustic data to underlying abstract linguistic messages, and acknowledged the importance to be attached to the fact that speech is produced by very specific physical mechanisms that are more readily accessible than neural

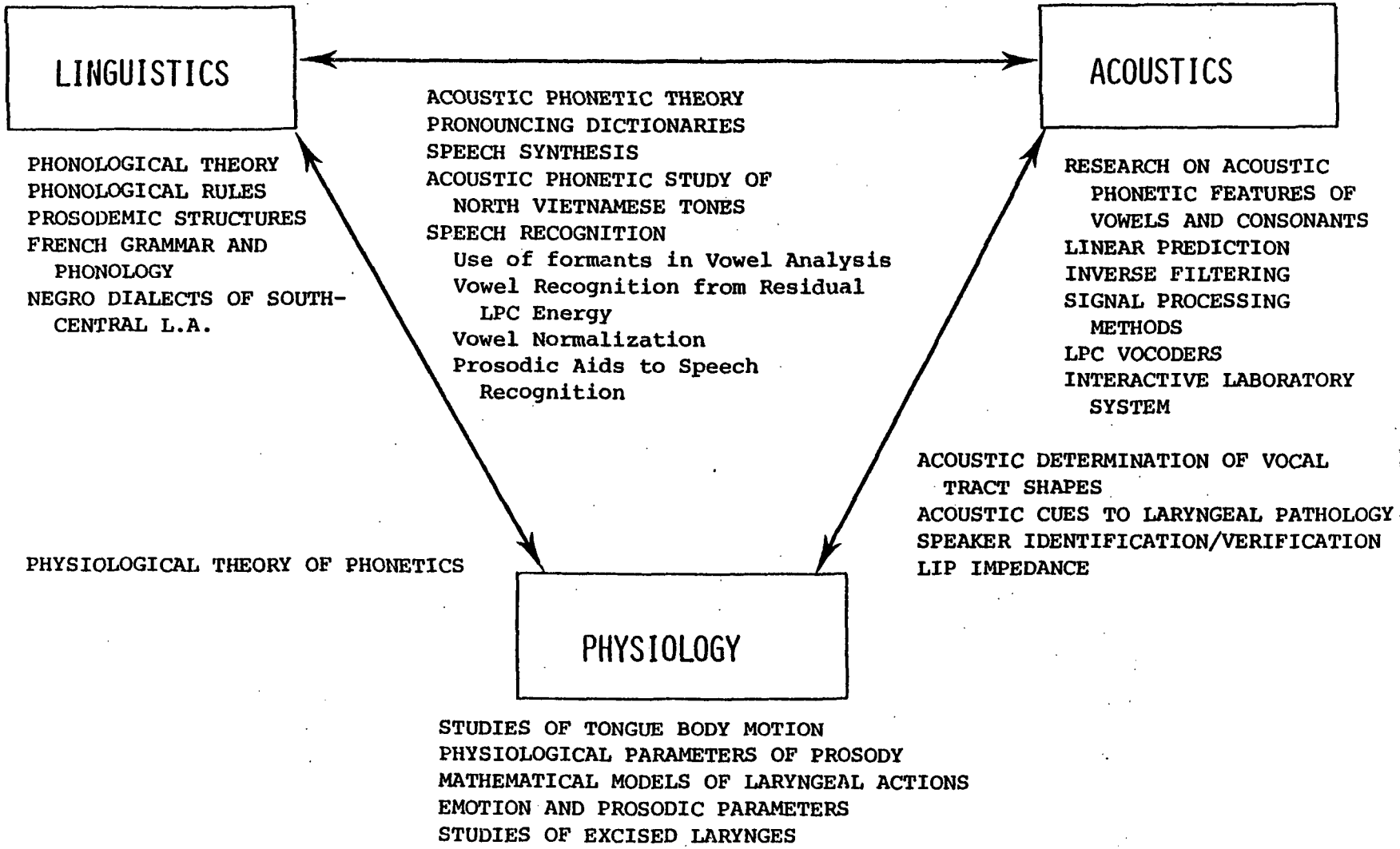

Figure 1. Speech Research Topics Studied in Previous Work at SCRL

commands of linguistic import. Consequently, physiology has played a major role in speech anlysis studies. In particular, it is frequently noted in speech recognition studies that manner of articulation (that is, whether a particular segment of speech is a vowel, a stop consonant, a fricative, a nasal consonant, or what) is more easily and reliably determined than place of articulation (such as, at the teeth, at the alveolar ridge, near the velum, etc.). Similarly, the physiological differences between male and female talkers is a notable reason for significant acoustic differences in their spoken vowels and consonants. The automatic recognition of voices is a way of linking acoustics to physiology. Two of the most impressive recent developments in speech science are concerned with (a) determining the vocal tract shape and (b) detecting laryngeal pathologies (such as cancer of the larynx), both directly from acoustic features. Major work in these areas was done at SCRL (Wakita, 1973, Davis, 1977).

While all this work impinges upon methods for speech recognition, there are some specific recognition projects that will be given special attention in the next subsection, to complete this review of previous (Pre-FY '78) work at SCRL.

3. Speech Recognition Studies at SCRL

Speech recognition research has been an important part of the projects and interests of the staff of SCRL since even before the founding of SCRL in 1966. In the late 1950's and early 1960's, while he was still with Bell Telephone Laboratories and the University of Michigan, Gordon Peterson outlined general models of automatic speech recognition and called for the use of linguistic structures, prosodics, and articulatory-based models to augment incoming acoustic information. Peterson was a leader in acoustic phonetic research and the author of works that are still among the most widely quoted in the field (e.g., Peterson and Barney, 1952). At the Univeristy of Michigan in 1963, he and Dr. June E. Shoup, the present Director of SCRL, conducted an epic-making course in Automatic Speech Recognition involving outstanding leaders in various related fields.

SCRL staff members have written several foundational papers concerning basic methods in speech recognition (Shoup; 1968, Broad, 1972 a,b; Broad and Shoup, 1975; Broad, 1976). In a frequently referenced paper, Broad (1972 a) described how to use formants in automatic speech recognition. Pilot experiments were also done on using residual energy of a linear prediction analysis to identify vowels. A method was developed for speech segmentation and normalization of spectral features based on the acoustically-derived vocal tract area functions (Kasuya and Wikita, 1976) and vocal tract length (Wakita, 1977). Automatic detection of syllabic nuclei was also studied at SCRL (Wakita and Kasuya, 1977).

The largest long-term effort in speech recognition at SCRL was undertaken within the ARPA SUR project. As a Support Contractor, SCRL developed new analysis tools and provided a variety of services for speech understanding system builders, such as:

- Doing a well-controlled phonemic analysis of a common database of "31 ARPA test sentences";
- Compiling lists of phonological rules;
- Developing methods for generating small dictionaries from lists of words related to a speech understanding task;
- Studying the feasibility of a common task for direct comparison of alternative speech understanding systems;
- Relating the literature on the location of syllable boundaries to the formal statements of phonological rules;
- Transcribing large speech databases orthographically, phonemically, and phonetically;
- Participating in planning meetings and workshops in acoustic parameterization, phonemic segmentation and labeling, and phonology.

SCRL cooperated with SDC, CMU, and BBN in their efforts to compile speech databases, develop and test segmentation and labeling schemes, and implement baseform dictionaries and phonological rules. My own work on prosodic aids to speech recognition, while initially done at Sperry Univac, may also now be considered part of the SCRL background in automatic speech recognition.

In summary of the SCRL work before FY '78, we have seen that general speech sciences work in linguistics, physiology, and acoustics, and the ties between those disciplines, have provided a general interdisciplinary background for a variety of specific studies in speech recognition. SCRL's specific ASR studies have ranged from detailed analysis and identificiation of vowels (using formants, residual LPC energy functions, and/or vocal tract area functions) to general theories of automatic speech recognition and rules for phonological anlysis. The pronouncing dictionary at SCRL is very large (300,000 entries), and orthographic, phonemic, and phonetic transcription methods are highly developed, and have been extensively used, at SCRL and by speech understanding system builders.

4. Tri-Services Contract to Review ARPA SUR and Survey the Current Technology

On July 20, 1977 SCRL was awarded a contract, sponsored by the Tri-Services and the Advanced Research Projects Agency, to review the five-year, $15-million ARPA SUR project and to survey the current technology in speech understanding. One task is to review and evaluate the performance of the speech understanding systems developed by Bolt Beranek and Newman (BBN), by the speech group at Carnegie Mellon University (CMU), and by the Systems Development Corporation (SDC, in cooperation with the Stanford Research Institute). We have read the various reports prepared by these groups, and have visited their laboratories to discuss the structures of their systems, the final performance results, their assessments of various aspects of their work, and their judgments about what work should now be done on speech understanding systems. We have concentrated on the techniques they consider to have been particularly successful, and have discussed with them the weakest points of their systems, and what further work is consequently needed. We have tried to understand why some systems have succeeded more than others, and have discussed what work these groups would want to do if given either one year or five years of further opportunity to extend their work. This provided us with a catalog of suggestions about work that deserves immediate attention, and work that should be included in the next major advance in speech understanding technology.

The significance of such a study can hardly be overemphasized. When ARPA initiated the ARPA SUR project over five years ago, the objective was to obtain a breakthrough in the ability of computers to understand spoken sentences. During two decades of prior research there had been repeated calls for overcoming the major hurdle separating moderately successful isolated-word-recognition systems from the unattained ideal of more natural uninterrupted voice communication with computers. Review articles had repeatedly called for the full use of language structures such as acoustic phonetics, coarticulation regularities, phonological rules, prosodic structures, syntax, and semantics (Lindgren, 1965; 1965; Hill, 1971; Lea, 1972; Broad, 1972 b). The ARPA project was the first large-scale effort to provide such a technology for understanding spoken sentences.

The original study report which formed the blueprint for the ARPA SUR project (Newell, et al., 1971) noted that successful speech understanding by computer depends on integrating various types of knowledge (e.g., acoustics, phonetics, syntax, etc.) and applying this multilevel information to the interpretation of utterances within a specific task domain. We are examining how ARPA SUR participants characterized these kinds of knowledge and organized these components into speech understanding systems, and are attempting to evaluate the various

components. The original ARPA SUR study group outlined goals that were very ambitious, given the fledgling state of continuous speech recognition and the defensive posture the field had following Pierce's (1969) pessimistic evaluation of speech recognition work (cf. Lea, 1970). Yet, the specific goals of the project are considered to have been substantially met by the HARPY speech understanding system that was demonstrated at Carnegie-Mellon University (CMU) on September 8, 1976. Other systems developed at BBN and SDC also attained some success in sentence understanding, though more ambitious goals of handling a sizeable subset of spoken English and conducting longer-range research appear to have prevented those systems from being tested, refined, and constrained adequately to attain the high (90%) semantic accuracy set down in the original goals. Still, many ideas and implementation techniques have been considered and tested in these systems that should be clearly understood, evaluated, and applied as appropriate in the development of future systems.

In addition to the CMU, BBN, and SDC systems, preliminary systems were developed at Lincoln Laboratory of MIT and Stanford Research Institute, and tested with some success in 1974. Also, supporting speech research efforts were conducted at Haskins Laboratories, Sperry Univac, and the University of California at Berkeley (transferred from the University of Michigan during the project), as well as at SCRL. We are also reviewing the scientific and technological advancements resulting from such work.

A five-year, $15-million, multiple-contractor program the size of the ARPA SUR project certainly deserves careful review and evaluation. Our responsibility as we see it is to evaluate the project with tomorrow in mind, not yesterday, so that we propose to address such questions as the following:

- What were the specific scientific and technological accomplishments in the SUR project?
- How has the state of the art in speech understanding advanced from 1971 to now?
- What problems in speech analysis became apparent from the efforts to provide systems that met the original specifications?
- What type of components produced the best results? The worst results? What are the sources of errors? In particular, what are the most common reasons for a system's being sidetracked into exploring wrong hypotheses about sentence structures?

Our review will hopefully provide an accurate picture of how the ARPA SUR project produced progressive steps in the technology of speech understanding systems. To complete a picture of the state of the art in 1977, we are attempting to relate the performance and techniques of the ARPA SUR systems to other work in the field. As soon as our ARPA SUR review is complete, we will study work at IBM, Sperry Univac, Bell Laboratories, ITT, Texas Instruments, Threshold Technology, and many other groups throughout the world. We hope to determine the adequacies and inadequacies of current capabilities and to help establish what is left to do to produce useful systems for a spectrum of applications. Some of the questions being addressed are:

- Where does the rest of the speech understanding field stand and how do the accomplishments of the ARPA/SUR program fit in with other work?

- What remains to be done to attain useful forms of speech understanding systems for DOD applications?

- How extendable are the current systems? Can they be made to operate with a natural ("habitable") subset of English? What is still needed to provide a spectrum of systems for handling various applications?

There are several dimensions of task difficulty in the speech understanding framework that need to be explored further. What happens to the performance of the alternative systems for speech understanding when:

- The language gets more complex and flexible

- The number of expected talkers increases

- Dialects and speech styles change

- The microphone or communication channel includes noise, bandwidth limitations, distortions, etc.

- The system cannot be as extensively trained (or not trained at all) for each talker

- The practical needs of real time operation on moderate-sized available computers are taken into full consideration

- Real task domains such as applications in the military services are tackled

- Very high accuracy in semantic understanding is demanded.

It is, of course, very difficult to assess the whole technology of speech understanding, and we have not been so presumptuous as to think we can answer all these (and other) questions by ourselves. We have distributed a questionnaire to about 100 researchers and technologists in speech recognition, seeking their opinions about the ARPA SUR project, the current technology, and the future work that is needed.

One of the primary goals of this Tri-Services study is to determine what needs to be done in future work on speech recognition and/or understanding. In addition to studies of all the documentation from the ARPA SUR project and other current work, and interactions with various workers to define the detailed adequacies and inadequacies of current systems and their components, we would like to work with ARPA and the military services to define what yet needs to be done and where to go from here. We all need the information being given at this workshop about DOD speech recognition applications, gaps in speech recognition capabilities, and possible programs for future development of useful systems.

5. Forthcoming Publications and Presentations

A primary outcome from the Tri-Services review and survey will be a series of publications summarizing what we have learned. The following is a list of publications and public presentations that are to appear:

- W. A. Lea and J. E. Shoup, Specific Contributions of the ARPA SUR Project to Speech Science, to be presented at the 94th Meeting of the Acoustical Society of America, Miami, Florida, December 14, 1977. (Abstract in J.A.S.A., vol. 62, Suppl. 1, Fall, 1977).

- W. A. Lea, President of a Special Session on "Speech Recognition: What is Needed Now?", International Phonetic Sciences Congress (IPS-77), Miami, Florida, December 19, 1977.

- J. E. Shoup, "Phonological Aspects of Speech Recognition:, to be presented at the IPS-77 Special Session on "Speech Recognition: What is Needed Now?", Miami, Florida, December 19, 1977.

- W. A. Lea and J. E. Shoup, "Gaps in the Technology of Speech Understanding", to appear in Proc. 1978 IEEE International

Conf. on Acoustics, Speech and Signal Processing, Tulsa, Oklahoma, April 10-12, 1978.

- TRENDS IN SPEECH RECOGNITION, a book edited by W. A. Lea, including the following papers by SCRL researchers:

  VOLUME I: (GENERAL ISSUES AND TRENDS)

  Ch. 1. The Value of Speech Recognition Systems (W.A. Lea)
  Ch. 4. Speech Understanding Systems: Past, Present and Future (W.A. Lea)
  Ch. 6. Phonological Aspects of Speech Recognition (J.E. Shoup)
  Ch. 7. Prosodic Aids to Speech Recognition (W.A. Lea)
  Ch. 17. Specific Contributions of ARPA SUR to Speech Science (W.A. Lea and J.E. Shoup)
  Ch. 23. Speech Recognition Work in Asia (H. Wakita and Shuzo Makino)
  Ch. 27. Speech Recognition: What is Needed Now? (W.A. Lea)

- W.A. Lea and J.E. Shoup to conduct a Workshop on Speech Understanding Technology and Its Applications, Washington D.C., Spring, 1978.

- W.A. Lea and J.E. Shoup, Review of the ARPA SUR Project and Survey of the Speech Understanding Field, Final Report on ONR Contract No. N00014-77-C-0570.

- W.A. Lea, "Advances in Speech Recognition", invited paper to appear in Proceedings of the IEEE, Special Issue on Pattern Recognition, May 1979.

- W.A. Lea, "Voice Input to Computers: An Overview", an invited talk to be presented at the National Computer Conference, Anaheim, CA, June 6-8, 1978.

Previous reviews of the ARPA SUR project have concentrated on final system performance and a general description of the systems developed. Our paper for the ASA meeting in Miami is intended to focus attention on the basic speech science results from the project. Only some of these results were actually incorporated into the final systems. Some were excluded in the final rush to complete work on operational but restricted systems, and some scientific contributions by the support contractors were not translated into specific algorithms for use in systems.

Dr. Shoup and I will endeavor to outline those gaps in speech understanding technology that need early attention, based on our survey of the current state of the art. Only some of these gaps can be included in the written version of the IEEE/ICASSP paper, which is due December 19, but more will be included in our oral presentation next April.

Also, in December, I am chairing a session at the IPS-77 Congress, which I have deliberately organized to focus international attention on the current technology and future needs in speech recognition. June E. Shoup is presenting an invited paper at that session on phonological aspects of recognition, which will be based on her review of phonological studies within ARPA SUR and the entire current technology.

The IPS-77 papers from that session, and 20 other papers from the most active groups throughout the USA and the world, will be included in a book which I am editing, and which is scheduled for publication in 1978. There is a section (composed of several papers) covering the ARPA SUR project, several papers on the need for speech recognition, tutorial papers about aspects of speech understanding system design, a series of papers about recent operational systems in the USA, and several survey articles dealing with the work in other countries. Much of our review and survey work is to be included in our chapters in that book. We have also been invited to provide a general review of the field for the Proceedings of the IEEE, a tutorial review for the IEEE Spectrum, and an overview for the National Computer Conference. Our final report will be issued next August, and will include all of our review and survey results, and our recommendations for future work.

6. Cataloging Available Services and Tools

Many computer programs have been developed in the course of the ARPA SUR project and other previous work. Extensive sets of sentences have been recorded, digitized, processed for important parameters, segmented and labeled with phonetic or phonemic category symbols. Some sentences have been transcribed by linguists, and in some cases those transcriptions have been time-locked to the speech waveform, so that valuable data for studying the acoustic phonetic, prosodic, and phonological structures of English sentences have been obtained. Also, valuable laboratory facilities have been developed for analyzing speech, playing it back (repeatedly, if desired, as in perception experiments), processing it for parameters, automatically segmentating and labeling, and many other speech-handling tasks. Statistical packages have been developed to keep track of such data, to automatically do analyses of regularities, and to plot such displays as histograms, discrimination thresholds, etc.

All this work should be cataloged and made available to all interested groups (where possible), so that duplication of efforts and costly diversions can be avoided in future studies. We hope to do some of that cataloging as time permits within our contract, and to outline general ways in which organizations like the IEEE Subcommittee on Speech Recognition can make such services and tools available to other researchers and developers of systems.

7. Future Plans

Obviously, since we are currently involved in a review and survey that will define what work should be undertaken in future studies, we cannot, and should not, at this time offer detailed plans for future work. We do have some general plans, and ideas for specific work that is in keeping with all that we have learned in our ARPA SUR review, discussions with other researchers, and survey to date. SCRL will continue to be involved in speech understanding studies, since the need for such facilities remains and there are significant gaps still to be filled in the available technology. In particular, we plan to pursue prosodics research and develop an improved and expanding capability in prosodic aids to speech understanding. Prosodics has been one of the knowledge sources that has been most obviously missing from previous systems, not only in our opinion but in the opinions of several other leading groups with whom we have visited (also, cf. Woods, 1974, p. 9; Wolf, 1977, p. 207).

Another major need reiterated by every group we conferred with is improved acoustic phonetic analysis (the so-called "front end" of many systems). SCRL has a long term history in such studies, and will presumably contribute to such work. However, the work in substantially improving acoustic phonetics aspects of recognition is very demanding and will require cooperative efforts by many different research, technology, and applications-oriented groups. It is particularly striking that major improvements in acoustic phonetics capabilities are needed despite decades of excellent work in that field, while ARPA's five year ambitious effort in artificial intelligence and higher level linguistics constraints has achieved such substantial progress that the bottleneck is again back in the difficult problem areas of acoustic segmentation, labeling and preliminaries to word identification.

I also see a definite need for future understanding systems to be tested on a common task (that is, evaluated with the same speech data and task domain) or else evaluated with carefully designed "performance metrics" that make it possible to decide whether 50% correct recognition on a difficult task is better or worse than 95% recognition on a much easier task. This is, for example, relevant in trying to comparatively evaluate the ARPA SUR systems developed at CMU, BBN, and SDC. Very little work has been done on performance metrics and task complexity metrics

that can make possible the comparative evaluation of alternative systems (cf. Goodman, 1976; Moore, 1977).

In conclusion, I have listed in Figure 1 the variety of research topics which SCRL has addressed in the past eleven years. I have sought to illustrate, with linear prediction and prosodic aids to speech understanding, some graphic examples of how interdisciplinary speech sciences research can readily lead to a variety of practical tools and provoke further scientific research. SCRL has conducted several studies in speech recognition, including providing transcription capabilities, prosodics research, and phonological analyses for the ARPA SUR project. We are currently engaged in a review of ARPA SUR, a survey of the speech understanding field, and a development of recommendations for future work in the field. We will be reporting our work in a number of publications, and already see several definite areas for further work, including prosodics, task complexity measurement (and performance metrics), and further advances in acoustic phonetic aspects of recognition.

8. References

D.J. Broad (1972 a), Basic Directions in Automatic Speech Recognition, Intern. J. Man-Machine Studies, vol. 4, 105-118.

D.J. Broad (1972 b), Formants in Automatic Speech Recognition, Intern. J. Man-Machine Studies, vol. 4, 411-424.

D.J. Broad (1976), Acoustic Discrimination Between (f) and (o) in a Single Speaker, Proc. 1976 IEEE Intern. Conf. on Acoustics, Speech, and Signal Processing, 162-165.

D.J. Broad and J.E. Shoup (1975), Concepts for Acoustic Phonetic Recognition. Speech Recognition, (D.R. Reddy, editor). New York: Academic Press, 243-274.

A. Chapanis (1975), Interactive Human Communication, Scientific American, March 1975, 36-42.

S.B. Davis (1977), Computer Evaluation of Laryngeal Pathology Based on Inverse Filtering of Speech, SCRL Monograph No. 13, SCRL, Santa Barbara, California.

P. Denes (1957), The Design and Operation of the Mechanical Speech Recognizer at University College London, The Journal of British Institution of Radio Engineers, vol. 19, 219-229.

R.G. Goodman (1976), Analysis of Languages for Man-Machine Voice Communication, Ph.D. Dissertation, Computer Sciences Dept., Carnegie-Mellon University, Pittsburgh, Pennsylvania.

D.R. Hill (1971), Man-Machine Interaction Using Speech, Advances in Computers, vol. 11, 165-230.

H. Kasuya and H. Wakita (1976), Speech Segmentation and Feature Normalization Based on Area Function, Proc. 1976 IEEE Intern. Conf. on Acoustics, Speech, and Signal Processing, 29-32.

W.A. Lea (1969), Evaluating Speech Recognition Work. J. Acoust. Soc. of America, Vol. 47, No. 6, 1612-1614 (L).

W.A. Lea (1972), Intonational Cues to the Constituent Structure and Phonemics of Spoken English, Ph.D. Dissertation School of Electrical Engineering, Purdue University, Lafayette, Indiana.

W.A. Lea (1976), Prosodic Aids to Speech Recognition: IX. Acoustic-Prosodic Patterns in Selected English Phrase Structures, Univac Report No. PX11963, Sperry Univac DSD, St. Paul, Minnesota.

W.A. Lea, M.F. Medress, and T.E. Skinner (1975), A Prosodically-Guided Speech Understanding Strategy, IEEE Trans. on Acoustics, Speech, and Signal Processing, vol. ASSP-23, 30-38.

N. Lindgren (1965), Machine Recognition of Human Language, IEEE Spectrum, vol. 2, March and April, 114-136 and 45-59.

J.D. Markel, Automatic Formant and Fundamental Frequency Extraction from a Digital Inverse Filter Formulation, 1972 Inter. Conf. on Speech Communication and Processing, Boston, Massachusetts, 81-84.

J.D. Markel and A.H. Gray, Jr. (1973), On Autocorrelation Equations with Application to Speech Analysis, IEEE Trans. on Audio and Electroacoustics, vol. AU-21, No. 2, 69-79.

J.D. Markel and A.H. Gray, Jr. (1974), A Linear Prediction Vocoder Simulation Based Upon the Autocorrelation Method, IEEE Trans. on Acoustics, Speech, and Signal Processing, vol. ASSP-22, No. 2, 124-134.

J.D. Markel and A.H. Gray, Jr. (1976), Linear Prediction on Speech. Berlin, Heidelberg, New York: Springer-Verlag.

R.K. Moore (1977), Evaluating Speech Recognizers, IEEE Trans. on Acoustics, Speech, and Signal Processing, vol. ASSP-25, No. 2, 178-183.

A. Newell, J. Barnett, J.W. Forgie, C.C. Green, D.H. Klatt, J.C.R. Licklider, J. Munson, D.R. Reddy, W.A. Woods, (1971), Speech Understanding Systems: Final Report of a Study Group, Computer Science Department, Carnegie-Mellon University, Pittsburgh, Pennsylvania.

G.E. Peterson and H. Barney (1952), Control Methods Used in a Study of the Vowels, J. Acoust. Soc. of America, vol. 24, 175-184.

G.E. Peterson and J.E. Shoup (1966), a Physiological Theory of Phonetics, Journal of Speech and Hearing Research, vol. 9, No. 1, 6-67. Also: The Elements of an Acoustic Phonetic Theory, Journal of Speech and Hearing Research, vol. 9, No. 1, 68-99.

J.R. Pierce (1969), Whither Speech Recognition? J. Acoust. Soc. of America, vol. 46, 1049-1051.

J.E. Shoup (1968), Approaches to Automatic Speech Recognition, Naval Reviews, June 1968, 11-17.

H. Wakita (1973), Direct Estimation of the Vocal Tract Shape by Inverse Filtering of Acoustic Speech Waveforms, IEEE Trans. on Audio and Electroacoustics, vol. AU-21, No. 5, 417-427.

H. Wakita (1977), Normalization of Vowels by Vocal-Tract Length and Its Application to Vowel Identification, accepted for publication in IEEE Trans. on Acoustics, Speech, and Signal Processing (in press).

H. Wakita and H. Kasuya (1977), A Study of Vowel Normalization and Identification in Connected Speech, Proc. 1977 Inter. Conf. on Acoustics, Speech, and Signal Processing, 648-651.

J. Wolf (1977), "Speech Recognition"; Invited Papers Presented at the 1974 IEEE Symposium (D.R. Reddy, Ed.). Book Review, in IEEE Trans. on Acoustics, Speech, and Signal Processing, vol. ASSP-25, No. 2, 207.

W.A. Woods (1974), Motivation and Overview of BBN SPEECHLIS: An Experimental Prototype for Speech Understanding Research, Proc. IEEE Symposium on Speech Recognition, Carnegie-Mellon University, Pittsburgh, Pennsylvania, 1-10.

9. Acknowledgements

This study is supported by the Tri-Services and the Advanced Research Projects Agency of the Department of Defense and is monitored by the Office of Naval Research under Contract #N00014-77-C-0570.

BIOGRAPHICAL SKETCH

Wayne A. Lea

Wayne A. Lea joined the staff of Speech Communications Research Laboratory, Inc. (SCRL) in August, 1977, and is serving as a Research Linguist and Research Engineer. His primary responsibility is as Co-Principal Investigator (with June E. Shoup), on a Tri-Services sponsored contract to review the five-year ARPA sponsored Speech Understanding Research project, to survey the current state of speech understanding technology, and to recommend further work relevant to future DOD applications for speech understanding systems. He also serves as Coordinator of Private Funding at SCRL.

Prior to joining SCRL, Dr. Lea was Co-Principal Investigator at Sperry Univac on a five-year research contract for ARPA, concerned with developing prosodic programs that located syllables, detected intonationally-marked phrase boundaries, and located stressed syllables. He also conducted a series of experiments on: human perception of stress patterns; prosodic correlates or linguistic structures; regularities in English rhythm, pause structures, accent, and intonations; and machine analysis of phrase boundaries, stress patterns, and phonetic structures. He proposed, and, for a time served as Co-Principal Investigator on, a project to spot occurrences of key words in continuous speech. Prior to his work at Sperry Univac, Dr. Lea has conducted prosodic research with the Purdue Research Foundation for two years. He also served four years with NASA at the Electronics research on speech recognition, mathematical linguistics, and the effectiveness of voice as a modality for man-machine interaction. Earlier he has worked on cybernetics, linguistics, and artificial intelligence projects at Montana State College and Massachusetts Institute of Technology.

Dr. Lea earned his Bachelor and Master of Science degrees in Electrical Engineering at Montana State College, then completed an Interdisciplinary Science Master's degree (with linguistics emphasis) and a professional Electrical Engineer degree at Massachusetts Institute of Technology. His doctorate is an interdisciplinary one (Electrical Engineering, English, and Audiology and Speech Sciences) from Prudue University. He has published over 50 reports, journal articles, conference papers, and book chapters, and is currently editing two books.