Journal of Research of the National Bureau of Standards
Volume 90, Number 5, September-October 1985

Performance Assessment of Automatic Speech Recognizers

David S. Pallett

National Bureau of Standards, Gaithersburg, MD 20899

Accepted: September 3, 1985

This paper discusses the factors known to influence the performance of automatic speech recognizers and describes test procedures for characterizing their performance. It is directed toward *all* the stakeholders in the speech community (researchers, vendors and users) consequently, the discussion of test procedures is not directed toward the needs of specific users to demonstrate the performance characteristics of any *one* specific algorithmic approach or particular product. It relies significantly on contributions from an emerging consensus standards activity, especially material developed within the IEEE Working Group on Speech I/O Performance Assessment.

Key words: acoustics; automation; computer technology; signal processing; speech recognition; speech understanding.

1. Summary

This paper identifies and documents factors influencing automated speech recognition performance. Procedures are outlined that are important in designing and implementing performance tests. Documentation is outlined which should clearly define test conditions. Definitions of terms are contained in the Appendix.

Definitive tests to *fully* characterize automatic speech recognizer or system performance cannot be specified at present. However, it is possible to design and conduct performance assessment tests that make use of widely available speech data bases, use test procedures similar to those used by others, and that are well documented. These tests provide valuable benchmark data and informative, though limited, predictive power. By contrast, tests that make use of speech data bases that are not made available to others and for which the test procedures and results are poorly documented provide little objective information on system performance. Such tests might be termed "incomparable" in that the data obtained cannot be meaningfully compared with data for other tests or for other systems.

Speech recognizers are the central element in speech recognition systems, and primary attention in this paper is directed to tests of recognizers as system components. Testing overall systems performance and the human-machine interface involves the more difficult task of developing measures of speech understanding. The factors described in this paper are necessary, but not sufficient concerns in tests of integrated human-machine and speech understanding systems.

A number of recommended testing procedures are described in general terms. These procedures are deliberately not specified in detail because it is recognized that no one detailed procedure could meet the widely-

About the Author: David S. Pallett is a physical scientist in the Systems Components Division of NBS' Institute for Computer Sciences and Technology.

varying needs for test data among researchers, vendors, and users. At present, decisions concerning the best way to implement specific tests and their applicability for research purposes, for commercial products and for proposed applications are best left to the judgment of the researcher, vendor, and user. It is, however, their responsibility to discuss and document the specific test procedures and data supporting their claims for algorithm or product performance. These discussions should be based on the considerations outlined herein.

Since automatic speech recognition is still an emerging technology, a standard terminology has not yet been established. Current activities with the IEEE Acoustics, Speech and Signal Processing Society include a Working Group on Speech I/O Systems Performance Assessment. This Working Group has contributed to the suggested definitions of terminology in this paper.

2. Introduction

Researchers and systems designers in a number of agencies of the Federal Government have worked closely to identify the capabilities of automatic speech recognition technology and to exchange research findings. Applications studied to date include data entry, package sorting, and command/control in aircraft cockpits. These studies have demonstrated the need for careful planning of trial applications and the value of thorough analysis of performance data. Industry is showing ever increasing interest in commercial applications of the technology.

However, until automatic speech recognition technology is a well established element in the human-machine interface, continuing efforts must be made to identify the relative importance of factors influencing performance and to develop and specify definitive test procedures. Tests conducted using these procedures will then serve to clearly demonstrate appropriate uses of the technology and to document the associated productivity benefits.

The paper is the first to report on the development of detailed and specific test procedures for performance assessment in the Institute for Computer Sciences and Technology at the National Bureau of Standards. The overall focus is assessing the performance of speech recognizers as system components, with emphasis on laboratory benchmark tests. The discussion in this first paper is introductory in nature. Continuing attention to these issues, along with the contributions of consensus standards groups, will result in the development of detailed procedures for both benchmark tests of speech recognizers and for measuring human performance.

3. Factors Which Influence Speech Recognizer Performance

Successful implementation of automatic speech recognition technology presents numerous challenges. In many cases these challenges are met through the selection and imposition of constraints on the many factors known to influence performance. Corresponding constraints must be imposed on the structure of performance tests if meaningful performance data are to be obtained. The need for these constraints arises, in large part, from the high inherent variability of unconstrained speech.

The inherent variability of speech arises from the nature of speech and the articulatory process. "Speech is based on a sequence of discrete sound segments that are linked in time. These segments, called phonemes, are assumed to have unique articulatory and acoustic characteristics. When speech sounds are connected to form larger linguistic units, the acoustic characteristics of a given phoneme will change as a function of its immediate phonetic environment because of the interaction among various anatomical structures (such as the tongue, lips, and vocal chords) and their different degrees of sluggishness [1][1]." This variability in the articulatory gestures involved in the production of speech and the interactions that arise from adjoining segments are important factors contributing to the difficulty in successfully implementing automatic recognition of continuous speech.

Humans have well developed abilities to adapt to and accommodate this variability, but at present it is a critical barrier to the automatic recognition of unconstrained speech. Automatic speech recognition systems have difficulty discriminating between linguistically meaningful and insignificant variations. This variability significantly complicates the process of testing.

There are numerous other factors that further complicate the task of successfully implementing and testing automatic speech recognition technology. These factors make it desirable to clearly anticipate the effects they may cause when designing, implementing, and documenting performance assessment tests. Appropriate recognition of these factors will increase the value of the test results as benchmarks for comparative purposes and enhance the predictive power of the tests. The following factors describe the main sources of variability that should be considered when testing automatic speech recognition technology.

[1] Figures in brackets indicate literature references.

Speech Related Factors

The form of the speech has a great effect on the difficulty of recognition. Isolated words or discrete utterances are easiest to recognize. Connected words, even if spoken carefully, are more difficult to recognize because the beginnings and ends of each word are affected by the adjacent words. Fluent or continuous speech is much more difficult to recognize becasue the sound segments (particularly those at the beginnings and ends of the words) tend to merge, and stress patterns affect the loudness and distinctiveness of vowels.

Speaker Related Factors

There are important differences in the way different individuals speak. Factors contributing to these differences which may affect performance and which can be readily documented include:

Age: Voice quality in adolescence and old age often differs from that in mid-life.

Sex: Certain speech characteristics, such as pitch and vocal tract length, tend to be gender-specific for adults. Some speech recognition systems employ features that may have been optimized for user groups such as adult males, so that it is important to document age and sex data for the test speaker population.

Dialect History: Some speakers are dialect chameleons, and adapt quickly and convincingly to the dialect characteristics of a new region. Others retain some qualities of previous dialects while changing other qualities. Because pronunciation of many words depends strongly on dialect, documentation of dialect data may be particularly important in tests of speaker-independent recognizers.

Speech Idiosyncrasies: Speech-associated anomalies can be expected to affect recognition performance adversely. Stuttered, lisped or slurred speech patterns and unusual characteristics should be identified and noted. For some individuals, speaker-generated noises such as lip smacks, tongue clicks, "um, ers, and ahs," etc., will degrade performance. Speech levels vary with changes in vocal effort and in the distance between the speaker's mouth and microphone.

Changes in rate of speech introduce additional complications. In words spoken rapidly, some of the sound segments may be shortened or deleted or altered in quality. A slow rate of speech may cause vowels to have drifting frequency spectra and extremely long silence gaps in plosive consonants (e.g., "p," "t," "d," etc.).

When enrolling, testing, and using speech recognizers, the use of chewing gum and smoking should be noted and/or controlled. Care should be taken when selecting a test speaker population so that these characteristics are appropriately represented.

Speech Variability: Individual speakers vary in the degree of consistency with which they repeat words. Some speakers produce nearly identical repetitions of individuals words or utterances, even under stressful conditions. Others produce highly varied repetitions (e.g., words such as "eight" with the final consonant occasionally deleted, or with highly variable pitch). The former category will make a recognizer perform best, and is sometimes referred to as "sheep." The latter is sometimes referred to as "goats."

Motivation and/or Fatigue: Degradation of performance for speaker-dependent systems can be expected as motivation degrades or fatigue increases. It is useful to obtain samples of speech under these conditions in order to estimate the degree of performance degradation.

Task Related Factors

The design of vocabularies for successful application of this technology is an important consideration. Limited size vocabularies require careful planning. Vocabularies should be natural to the task and sufficiently distinct to ensure recognition with few substitution errors.

Performance is greatly improved by the imposition of syntactical constraints. In many task dialogs there are only a few possible choices at each point in the task, thus making the recognition task much simpler, faster, and more reliable.

Physical exertion, fatigue, and other stressing factors must be considered and documented in designing experiments and assessing performance. The voice pitch and loudness or vocal effort of the speaker change due to stress, as do the spectral components.

Environmental Factors

The input speech signal to a speech recognizer is affected by background noise, reverberation, and transmission channel phenomena (e.g., the use of telephone lines or wireless microphones). These environmental factors may lead to spurious responses by the recognizer. The performance of speech recognizers will generally be lower when telephone lines are used for input than with direct microphone input because the frequency response is limited and noise artifacts make correct recognition more difficult. The use of wireless microphones may lead to recognition errors due to transmission channel cross-talk, RF interference, signal fading, dropouts, etc.

Other Factors

Human recognition of speech involves an imperfectly known set of decision criteria. Automatic speech recognition devices apply specific, but (to some degree) arbitrarily chosen, decision criteria in order to effect recognition. Optimum settings of these decision criteria, including the associated reject thresholds, are extremely important. However, the optimum settings of these decision criteria are controlled by the vocabulary, the design of applications software (i.e., the implementation of syntactic constraints, error-correction protocols, etc.), and the characteristics of the individual user's speech and personal preference. Experimentation is required in order to determine the optimum setting of the reject thresholds. As an alternative to the selection and use of optimum settings of the reject thresholds, the reject capability may be disabled, to simulate a forced choice response. This procedure is frequently chosen for benchmark tests.

In some cases, the system may also have the ability to return ordered word lists. Typically, these word lists are ordered according to the distance measure between the input word and the reference templates or word models or in order of descending probability. The application of higher level constraints such as syntax then may lead to correct identification of the utterance. While this process may emulate human decision criteria and typical decision trees, it can complicate assessment.

4. Considerations in Developing Test Procedures

The design and implementation of tests to define the performance of automatic speech recognizers requires that attention be paid to many of the previously described factors influencing performance. A systematic process of experimental design and testing is indicated in this section to account for these factors. This process includes:

- Selecting an experimental design that either
  (a) models an application, or
  (b) provides benchmark data.
- Selecting speakers to represent the user population or some relevant subset.
- Selecting a test vocabulary that either
  (a) exemplifies that used in an application, or
  (b) has been used by others for benchmark test purposes.
- Training the system, or constructing the reference patterns to be used by speaker-dependent recognizers.
- Characterizing the test environment in order to document complicating factors such as factory noise, communications channel limitations, or task-related factors.
- Recording the test material to permit verification of the validity of the test results and reuse of the test material.
- Scoring the test results. Procedures are outlined for both isolated and connected word data.
- Pragmatic considerations to ensure that equipment is properly operating, that tests are conducted in a manner that is consistent with manufacturer's recommendations, and other related factors.
- Statistical considerations to indicate the statistical validity of performance data.
- Documentation of test conditions and performance data to allow evaluation of published data.

Tests designed and carried out accounting for these factors will be valuable in identifying the strengths and weaknesses of automatic speech recognition systems. The importance of performance assessment procedures has been emphasized in a recent study by the Committee on Computerized Speech Recognition Technologies of the National Research Council. Their report [2] recommends that: "... performance should be measured within a realistic task scenario, both within the laboratory and in actual operational settings, including worst case conditions. Laboratory benchmark tests using standard vocabularies, experienced users, and controlled environments are useful for comparing recognizers, but they are not efficient for predicting actual performance in operational systems. Adequate methods are needed for measuring both human and recognizer performance under realistic conditions. The importance of performance measurement techniques cannot be over emphasized since they provide the data for decisions about system design and effectiveness ...."

Experimental Design

There are two complementary approaches to designing performance assessment tests. These approaches are summarized in table 1.

In one approach, a set of benchmark test conditions is defined (e.g., use of a "standard" speech vocabulary and data base, and no use of syntax to actively control the recognition vocabulary). Little or no effort is taken to model an application. This approach provides valuable comparative performance information. It does not directly predict performance in real applications.

Table 1. Alternative approaches to test design.

Test Conditions	*Benchmark Tests*	*Applications Tests*
Vocabulary	Benchmark or Reference Vocabulary	Applications Specific (Task) Vocabulary
Data Base	Widely Available Recorded Data Base	(Variable)
Use of Syntax	Little or No Use of Syntax	Syntactically Constrained Word Sequence or Imposed Task Grammar
User Interaction	None	(Variable)
Predictive Power	Very Limited	Less Limited
Data Analysis	Detailed	(Variable)
Documentation	Thorough	(Variable)

A second approach consists of carefully selecting test conditions in order to simulate a field application. The use of syntactically constrained word sequences may dramatically enhance performance and is acceptable for user applications. The design of a test vocabulary should include specifying the structure of the grammar and the frequency of occurrence of each item. This approach may have greater predictive power in inferring performance in specific applications, but it complicates comparisons between differing applications. Because of the diverse applications proposed for recognizers, simulation of many different applications and the needs of differing users becomes very difficult and/or costly.

In both approaches to testing, simple averages such as error rates, recognition accuracy, etc., are often inadequate to indicate performance. It is important to determine and document the most frequently occurring confusion pairs (e.g., "five-nine" confusions, where a spoken five is recognized incorrectly as "nine"). Presentation of this data in the form of a confusion matrix is very informative. For an N-word vocabulary, the confusion matrix is an N-by-N-matrix of input versus output, a form of stimulus-response matrix representation. Correct recognition responses fall along the diagonal of this matrix, and substitution responses comprise the off-diagonal elements.

Selecting the Test Speaker Population

In both benchmark tests and in applications tests, care should be taken to select speakers for the tests that are in some sense representative of the ultimate users of the technology. For example, in applications tests of industrial quality control data entry systems, the most valuable test speakers will ordinarily be quality control personnel. In research or benchmark tests, the test speakers are ordinary adult males and/or females with "neutral" dialects. In extraordinary circumstances, efforts are taken to obtain representative speakers with regional dialects. However, representative sampling of all potential users is not always possible or necessary. The characteristics of the test speakers and both their user training and system enrollment procedures should be documented. While all of the documented factors may not significantly affect performance, the documentation will indicate to others whether the test group is of particular interest or relevance.

Some recognizers impose limitations on the duration of words, or of silence gaps within words considered as single words or strings. Other constraints may apply to the number of words which may constitute a connected string. These constraints may have important consequences in some applications and for some individual speakers (e.g., if the durations of the speaker's stop gaps are longer than a limit set by the manufacturer, the word or phrase may be segmented into two utterances, and will not be correctly recognized).

Because speech recognizers use enrollment data to build reference template sets, prototypes, or other internal representations of the words to be recognized, it is important that the enrollment data for speaker-dependent systems provide representative samples of the user's speech. The enrollment and test data should include speech that is characteristic of the application, possibly including fatigued or stressed speech. These requirements may complicate enrollment and test procedures and, when slighted, generally result in lower performance in an application. Other important factors include the degree of cooperation of the users and their familiarity with the equipment.

Automatic speech recognition algorithms and commercial systems perform best on systems trained for the intended user's voice. Such speaker dependent recognizers provide some degree of language independence, depending on the type of acoustic-phonetic representation or pattern matching algorithm used by the device. They may perform equally well when used with several languages. However, speaker independent recognizers are expected to be language and dialect dependent to the extent that they rely on phonological rules and specific data bases for the development of internal representations. The issue of language or dialect independence may be very important for some applications.

Speaker independent systems do not rely on the data obtained from the individual user's voice. Rather, they are designed using training or enrollment data from many speakers and incorporate representations (e.g., template sets derived by studying clusters of individual speakers' templates or word models derived from statistical analysis of many individual speakers' word models) based on features which are presumed not to vary from individual to individual. This is a crucial assumption (that the system relies on features that are relatively consistent) and its successful implementation is the key to success in speaker independent automatic speech recognition technology. It is essential to ensure that the most important variabilities and dialect related factors have been accounted for when designing and testing such systems. These requirements become increasingly challenging if large vocabularies are required and response must be available ina time period comparable with the duration of an utterance (i.e., real-time recognition).

When selecting test speakers for "speaker independent" systems there are a number of special concerns. Perhaps most importantly, a representative sampling of the intended user population should be obtained in order to appropriately represent regional dialect and/or transmission channel effects for the intended user and applications population. A statement describing the efforts taken to represent the user population should be included as part of the documentation. When conducting tests of these systems, it is important to exclude data from the test material that might have been used in constructing internal representations used by the recognizer. Casual recognition experiments using template sets generated from one person or a small number of people typically demonstrate highly variable performance. Sometimes recognition performance may be quite good or quite poor for some individuals, and frequently there will be good performance on some words and poor performance on others. For these reasons, casual experimentation to demonstrate "speaker independence" for systems designed to be speaker dependent is not recommended.

Selecting the Test Vocabulary

The actual performance of any given speech recognition system in both benchmark tests and applications is critically dependent upon the vocabulary items that must be distinguished at any given time. Both the number of items to be distinguished and the acoustic similarity or complexity of these items are critical factors.

Brief monosyllabic words (e.g., yes, no, go, the natural alphabet except for "w" etc.) are more difficult to recognize than longer polysyllabic words or brief phrases spoken and intended to be recognized as single items (e.g., Massachusetts, California, "start printing," "left bracket"). These more complex utterances contain much more acoustic information and redundancy than monosyllables. In actual applications, this fact is used to construct vocabularies that retain many of the qualities of a natural interaction while selecting somewhat more complex acoustical characteristics to maximize system performance.

For these reasons, it is necessary to explicitly state the test vocabulary. It is, of course, desirable to use a test vocabulary that is identical to the intended vocabulary for the application.

One parameter often used to characterize recognition system performance is that of the vocabulary size. Vocabulary sizes, ranging from approximately 40 to several hundred words, are not unusal at present. However, in order to enhance performance, it is often appropriate to use syntactic constraints. This is implemented through the imposition of an artificial language grammar to constrain the vocabulary choices at each stage of a task in a given application. In many cases, this is not only appropriate but will lead to significantly enhanced productivity by imposing a desired order for completing the intended task.

Restricted vocabularies and formatted messages are widely used for speech communications in situations such as air traffic control and military tasks in which high speech comprehension is required. Acceptance of these constraints in isolated and connected word speech recognition applications will result in higher performance, but must be explicitly stated when documenting system performance.

It is important to distinguish between the total vocabulary capacity (typically a function of total memory available to the system) and other measures of the effective vocabulary size (typically functions of the structure of the imposed artificial language grammar). For artificially constrained tasks, the average number of alternative words that the system has to choose from at any time is given by the "perplexity," or dynamic branching factor for the imposed artificial language.

A 10-word recognizer, requiring discrimination between the digits (0–9) with all transitions equally likely is typically more difficult than a system with a several hundred word total vocabulary and branching factor of only 5. Even if the larger vocabulary has a branching factor of 10, the larger vocabulary may be easier than the 10-word digit vocabulary if the vocabulary words tend to be longer and more discriminable than the digits (eight of which are monosyllables).

The benefits achieved through the use of syntactic constraints may be better addressed in separately documented tests. In syntactically constrained tasks, performance results ought to be reported with the following information to describe the characteristics of the imposed grammar.

(a) Complete description of the task grammar including full specification of the vocabulary at each task state or menu choice.
(b) Frequencies of transitions from each task state to successive states.
(c) Dynamic branching factor or perplexity.
(d) Frequency of occurrence in the test material of each vocabulary item.

Training

Two meanings of the word "training" are sometimes found in the literature of current speech recognition technology. A clear distinction must be made between them.

In one meaning, the user's speech is used to "train" the recognizer for the specific test or applications vocabulary. During this process, reference patterns ("template sets," "voice patterns," "voice prints," etc.) or more complex word models are developed and become the stored internal representations used for comparison with subsequently input speech in the recognition process. This process is referred to as "enrollment" without ambiguity.

A second meaning of the term "training" refers to that process in which the user of a recognizer becomes familiar with the device or system. During this "user training," many factors may combine to influence the user's speech. Generally, familiarization with the devices leads to improved performance, and the user learns to adapt to explicit, as well as implicit constraints on the form of the input speech.

One factor in user training that tends to improve performance uses feedback provided to the user. To date, most recorded speech data base material has not been obtained under circumstances allowing user feedback. The recorded speech data base material has been obtained in response to prompts or in list-reading tasks. The nature of the feedback provided to the test speaker should be documented along with a description of any prompts provided to the user or the tasks conducted by the user while providing test material.

In tests conducted on integrated systems (as opposed to tests on system components), time must be allowed for familiarization with the system and to observe the nature of performance improvement or degradation. In most cases, after a period of initial user training, performance can be improved significantly by simply re-enrolling the user. The new internal representations should then be more representative of the experienced user's typical speech, and poor initial performance due to the lack of user familiarity will be improved. Documented performance ought to represent the data obtained with experienced or fully trained users.

Characterizing the Environment

Both the operational environment and the speech signal transmission system providing input to speech recognition systems are important environmental factors influencing performance. For example, in an industrial quality control voice data entry application, the talker's environment might be a noisy factory floor, while the speech signal transmission environment may be a wireless microphone. When modeling an application, the acoustic environment and signal transmission channel should closely simulate the intended operational environment.

When access to the actual intended operational environment is limited or costly (e.g., in tests of systems for use in operational aircraft), using accurate simulations can provide a cost-effective test environment. By accurately modelling the environment, the value of such tests is enhanced by increasing the correlation between the test data obtained in the simulation and the actual operational environment.

Because laboratory test data are often not applicable to the user's operational environment, the responsibility for tests in operational environments becomes a critical element in dialogues between vendors and users.

When an actual operational environment is used, as for laboratory tests, care must be taken to control and document all potentially relevant characteristics of the test environment. Environmental noise tends to interfere with communication between humans. It also tends to degrade speech recognition system performance and it is best to separately conduct certain benchmark tests in which all background and transmission noise is minimized. These tests tend to provide information on optimum system performance because acoustic-phonetic information is not obscured by the noise. Comparison of benchmark test data with operational data can indicate the existence of noise-related limitations on performance for which noise control measures or improved transmission channels can lead to improved performance.

There are at least three types of noise that can affect performance:

-Ambient or background noise. This noise originates with the operation of nearby machinery such as office equipment, with ventilation systems, with people conversing in the vicinity of the user, or within the applications environment such as the crewspace of an aircraft. When microphones are located some distance from the talker's mouth, reflections from nearby surfaces such as desk-tops and room walls constitute a form of multipath interference that can be comparable to increased ambient noise in degrading system performance.

-Transmission or channel noise. Such noise is inherent in using wireless microphones or telephone lines, and (for long distance lines, in particular), may be due to signal processing devices such as echo suppression, multiplex, or satellite transmission systems.

-Inadvertent test speaker noises. These may originate in coughs, stammers, "ers," "ums," excessive breath noise, and speech extraneous to the selected recognition vocabulary.

Characterization of noise is important in interpreting operational test results. Attention should be directed to performance limitations that may be due to the following factors:

1) *The noise experienced by the speaker*. Speakers may modify their speech significantly in the presence of high noise. Typical modifications include speaking more loudly or slowly and taking care to articulate more carefully than otherwise. Minimal characterization provides the A-weighted sound level (dBA) experienced by the user. Because masks, helmets, and some headsets affect the perceived noise level, their use by the test speaker should be noted. More detailed characterization should include spectral content (e.g., third-octave band analyses) and the temporal nature (e.g., steady-state, intermittent or impulsive). Impulsive noise can lead to substantial degradations in performance, but full characterization of this noise is difficult to achieve without sophisticated instrumentation.

2) *The speech signal-to-noise input to the recognizer (prior to any recognition system signal processing)*. The use of a noise-cancelling microphone can effectively eliminate much of the noise environment of the test speaker, even in a high noise environment. However, the signal-to-noise properties of the signal input to the recognizer may be a critical factor in limiting performance in noisy environments. The relative importance of the differing characteristics of speech emitted in a noisy environment vs. the degraded signal-to-noise properties is not yet well understood. Different algorithms and/or devices are probably affected to differing degrees.

3) *Type and characteristics of the microphone*. Useful characteristics to note include close-talking or noise-cancelling, directionality, whether push-to talk or otherwise manually switched, distance from the speaker's mouth, etc. The effectiveness and frequency response of noise cancelling microphones are influenced by the distance to the sound sources. Thus, placement of the microphones should be documented.

4) *Verbal characterization and description of the origin of the noise*. Typical characterizations use terms such as "buzzy," "hum," "static," etc., and descriptions of the origin are "office environment," "package sorting machinery," "receiving platform," etc.

If a speech signal transmission system other than direct microphone input is used, attention should be directed to these additional factors:

5) *Limitations on the transmission channel bandwidth and frequency response*. Cite the upper and lower cut-off frequencies and any significant deviations from flat frequency response over the cited bandwidth.

6) *Other limitations on the transmission channel*. Significant performance limitations may be due to other effects such as automatic volume control attack and release characteristics, signal compression and/or limiting, phase distortion, additive noise in transmission, etc. In general, these effects are difficult to characterize.

While it is possible and, in many cases, desirable to record speakers in sound isolated (low ambient noise) and anechoic (dead) environments in order to access subtle details of the speech signal, care must be taken in generalizing these observations to infer the nature of speech in the presence of noise. A first-order procedure involves the addition of white, pink, or carefully shaped noise spectra to the speech data after they are collected. Factors associated with the acoustic environment that influence the speakers include both ambient noise level and the degree of reverberation.

Speakers may compensate for these factors by speaking more loudly or enunciating more clearly or slowly. It should be specified what ambient noise was audible to the speaker at the time of collection.

In addition to acoustic environmental influences on the speaker, the task environment modifies the speaker's performance. Different types of tasks will affect the speaker's speech to varying degrees. Routine speech tasks such as list-reading or responding to visual prompts displayed on terminals produce less word-to-word variations than speech produced when there are concurrent physical tasks or shifts in attentional focus usually associated with other cognitive activities such as inspection, measurement, etc. For these reasons, higher recognition system performance can be expected from speech obtained from list reading than when there is concurrent tasking, and the talker's task environment should be fully described in the results.

Recording the Test Material

Many performance assessment tests make use of recorded speech data bases. Others are conducted "live." The general practice of recording the test material, even for "live" testing, is recommended. The recorded material provides a means of replicating the results obtained and verifying that the test material was properly input to the system. It also provides material to be used in additional measurements on similar systems as well as for analysis of the input audio signal. Using recorded test material offer the advantage of providing samples of speech obtained on different occasions, separated by days or weeks. This can account for some of the day-to-day variation and may more accurately model potential applications.

Recorded speech data bases exist in both digitally recorded and stored formats and analog recorded formats. The digital formats have greater signal-to-noise ratio than the analog format. Reference data bases that are widely used in research and testing have been recorded with 16-bit samples at sample rates from 10.0 to 20.0 Khz. Signal-to-noise ratios in excess of 90 dB are feasible using this technology. A widely used format for analog recordings is the use of quarter-inch magnetic tape at 7.5 inches per second, providing maximum signal-to-noise ratios of the order of 60 dB. The use of cassette tape recorders is not generally recommended for benchmark test purposes for a number of reasons including increased print-through.

Another advantage of using digital storage for data base material is that each speech token may readily be assigned an accompanying "header" to indicate the origin of that particular token. Comparable systems are feasible using analog storage, storing the header information in an encoded analog signal on the second channel of a two-channel tape recorder, but these systems may require specialized interfaces.

Newly developed recording technology includes use of 14 or 16 bit digital sampling and pulse code modulation systems to encode the signals for storage on Beta or VHS format video recorders, referred to as PCM/VCR recording technology. This technology offers many of the advantages of digital sampling and storage at lower costs for the storage medium than more traditional digital storage media, and offers the capability of copying the data with less degradation than for analog recordings.

A number of speech data bases have been widely used in testing and serve to provide test material for benchmarks. They are available in several recorded formats [3].

Scoring Isolated Word Data

Scoring isolated word recognition systems performance presents fewer challenges than for connected word systems. The relative ease in scoring isolated word data arises from the fact that most errors tend to be substitutions: deletions or insertions are easily identified when they do occur.

It is ordinarily presumed that the individual speech tokens (e.g., words or short phrases with minimal intraword pauses) are separated in time by pauses that are long enough to permit the recognition system to respond. Indications that this may not be the case will be found if there is a high incidence of deletions or substitutions, and it should be noted that the origin of these errors may be due to a problem with the system's response time for the data base used for these tests.

Prior to detailed data analysis, it is instructive to critically listen to the recorded test material, particularly for those portions of the test material where unusual numbers of errors may have occurred. These errors may be due to noise artifacts or departures from proper script-reading or responses to prompts. If this is the case, the recorded tokens or artifacts must be editorially deleted from the test material prior to testing. Objective analysis of the data must include full documentation of these decisions regarding certification of the test material. It is preferable not to delete any data if the process of obtaining and using the test material was carefully structured and monitored.

Preliminary analysis of the performance data should identify and tabulate words which were correctly recognized, words provided as input for which substitution errors occurred, words provided as input for which there was no response (deletion errors or rejections), and instances in which a response occurred without a corresponding appropriate input (insertion errors). The raw data should be summarized by determining the corresponding correct recognition percent as well as the substitution, deletion, and insertion error percent.

In comparative testing of differing systems, it is generally preferable to disable the reject capability, so that each system returns a forced choice response. In this case, words provided as input for which there is no response are unambiguously classified as leading to deletion errors.

In performing benchmark tests to compare different recognizers, the removal of all syntax constraints may be preferable. These constraints, like the setting of reject threshold, may affect systems differently. If, following data analysis, recognition errors are concentrated on several specific words or utterances, then re-enrolling the speaker on these words or substituting acoustically distinctive synonymous words may substantially improve performance.

In other tests, particularly those at the integrated system level or in modelling an application, the use of the reject capability is an important feature that should be included in the test program. Tests in this case should document the settings of the reject threshold and/or other decision criteria, identify and tabulate words input to the system for which the reject response occurred, and determine the rejection percent. It is also valuable to determine and document the ratio of total errors to rejections, because this information may be useful in the design of applications software.

More detailed analysis of systems performance can be documented and easily reviewed by constructing a confusion matrix. Analysis of these data will provide valuable insights into systems performance and the design of successful vocabularies.

Another useful measure may be appropriate for those systems that present several ranked words for approval. In this case, recognition accuracy as a function of the word rank is a useful parameter, since it provides a measure of the probability that the second, third, ... Nth candidate is correct if the higher ranked candidate is incorrect. If it is known *a priori* that a recognizer will be implemented in an application that will impose higher level constraints, such as a syntactically controlled (sub-) vocabulary, then it is appropriate to determine and report the probabilities that the correct word is to be found among the top N candidates on the ordered list. This practice is inappropriate if the imposition of higher level constraints is impractical in a typical application.

It is sometimes desirable to have measures of a system's capacity to reject words that are not in its recognition vocabulary. This is particularly appropriate for those applications involving inexperienced users or those unaccustomed to using artificial grammars or syntactic constraints. For experienced users, it may be safe to assume that the input is limited to words in the recognition vocabulary, in which case out-of-vocabulary rejection capabilities are less critical.

In order to test a system's capability to reject out-of-vocabulary utterances a secondary test can be performed using the same recognition data base used for other tests. In this case, however, a subset of the recognition vocabulary is selected and the system is reenrolled using only this subset of the entire test vocabulary. The entire data base is then used for test purposes, with responses that occur for words that are not part of the active vocabulary (the selected subset) being classified as "false acceptances." Documentation in such a test must include the total test vocabulary and the specified active vocabulary.

Scoring Connected Word Data

There are several ways to score recognition performance on connected word strings. The most stringent method is to record the percentage of strings completely recognized, i.e., the number of correct strings divided by the total number of strings tested. Another method is to calculate the percentage of individual words correctly recognized. The number of substitutions, deletions and insertions is calculated for each string, and the total error count is divided by the total number of words in the strings tested. The string scoring may be done by procedures involving strict left-to-right alignment or by a best-case pattern match.

Left-to-right alignment procedures involve matching each word in the input string with a corresponding word in the response string starting with the first (leftmost) member of each string. Obviously, the occurrence of an insertion or deletion will shift the position of words in the response string so that succeeding responses will be compared with inappropriate members of the string. For example, if the input string is 12345, and a deletion error results in a response string 1345, a left-to-right alignment procedure correctly scores the first digit as a correct recognition, but would cite the three following responses as substitution errors (e.g., "3" for "2," "4" for "3," "5" for "4") and would detect the presence of a deletion error only at the last digit (e.g., no response for "5"). In this case, one correct recognition, three substitutions, and one deletion would be indicated where in fact there were four correct recognitions, and one deletion (and the deletion error is, in fact, mis-identified). This left-to-right pattern match procedure, though well defined and easy to implement, is in many cases a very poor or worst-case pattern match.

When using best-case pattern match procedures, individual words are matched so as to minimize the number of errors within the string. That is, if the input string is 12345, and the output string 1345, it is inferred that there were four correct recognitions and one deletion.

Selection of the most appropriate scoring method involves consideration of the relevant application, and particularly the manner of verification and correction by the speaker. Where the manner of correction involves repetition of an entire string, the string error rate may be most appropriate. The aligned word recognition scores would be appropriate measures for those cases in which correction may be possible by backing up one word at a time for the end of the string.

Using best-case pattern match procedures tends to avoid some of these complications, but there are no generally agreed-upon procedures to uniquely define the best-case match criteria. Specific details of these procedures are beyond the scope of this paper. Purchasers of systems for which these considerations are significant should discuss the scoring procedures used by vendors.

Pragmatic Considerations

Prior to conducting tests, care should be taken to make sure that the equipment is functioning properly. Because recognition systems are designed to perform with distorted and/or variable input, determining proper functioning is not a simple task. Malfunctioning components can masquerade as an imperceptible input distortion, and the system will appear to work but not as well as it should. Check procedures should include tests to confirm consistency of recognition results with results obtained previously using recorded speech, checks of input amplitude settings, and the use of any available software diagnostics. These tests should be routinely conducted at the time of testing.

In tests of commercially available systems, the manufacturer's recommendations should be followed in order to obtain optimum performance. If the manufacturer's recommendations are not followed, some degradation in performance may be expected.

Manufacturers may suggest procedures to use regarding:

-Recommended number of enrollment tokens.

-Presentation order of enrollment tokens.

-Required minimal interval pause duration.

-Amplitude and gain control settings (to accurately simulate live input if recorded input is used). Amplitude settings should not be readjusted, once set.

-Microphone position.

Specify if a "press to talk" switch is used. The use of "press to talk" microphones provides an input signal to the recognizer that has very little or no signal amplitude between words. Depending on the particular recognizer's procedure for accommodating input signals during inter-word pauses, this may lead to either improved or degraded performance, relative to the use of conventional unswitched microphones. Choice of "open" or "press-to-talk" microphones should be determined by operational considerations (e.g., if it is an accepted practice in a proposed application or if it is required to activate a remote system) as well as whether the use of one or the other may lead to optimum performance with a given recognizer.

Proper connection of peripheral devices and electronic components should be checked before testing to eliminate ground loops and extraneous noise. The speech (audio) signal input to the recognizer should be monitored by the experimenter to verify that no extraneous noise is being introduced.

Statistical Considerations

Performance assessment tests of any automatic speech recognition system require that a large number of speech tokens be input to the system by many users and that detailed analysis of the test data be conducted. Thorough testing requires the use of large test speech data bases, substantial data storage, and time. Disregard for these facts inevitably leads to misleading conclusions regarding system performance.

Researchers, vendors, users and developers of automatic speech recognition technology each have different needs for performance data. The significance and interpretation of the data vary because of the different goals each group seeks to achieve. Consequently the degree of concern for the statistical validity of the performance data is variable.

Factors that need to be considered in structuring statistically valid performance tests include the size of the test speaker (user) group, the number of test tokens, and the amount of enrollment material provided to the system.

For benchmark testing, a concise statement of the number of test speakers, number of test utterances, and number of errors of each type should be given. It is recommended that the performance documentation should include a statement of the total error rate and the confidence level implied by each statistic. Statistical tables should be consulted to interpret the results, and the assumptions made in computing the statistics should be stated explicitly [4].

For tests that model an application, statistically based considerations include sampling the intended user population and range of tasks to be implemented using speech recognition and defining the variability in the noise environment.

For speaker independent recognition technology, particular attention need be paid to sampling the intended user population and communications channels. Dialect-related effects and variations in the quality of telephone connections make it difficult to obtain consistent performance from current low-cost remote access speaker independent recognition technology. Testing of this technology must be based on large speech data bases.

To obtain optimal performance from each of the systems to be compared in a benchmark test the appropriate vendor-recommended training procedure must be followed. Because some recognizers make use of single-token enrollment, while others build increasingly more reliable statistically based word models in the process of enrollment (and, possibly, in operating in a speaker-adaptive mode), appropriate enrollment procedures often vary significantly from one system to another.

No generally accepted rules have yet been developed for statistically reliable speech recognition system test procedures. In view of the many factors influencing performance, most researchers and vendors attempt to carefully control known sources of variable performance. Those researchers and vendors whose products build increasing accuracy with statistically large enrollment data are particularly conscious of the need for statistically large enrollment and test data bases. Though no generally accepted rules for adequate statistical sampling presently exist, data analysis should seek to define the distribution of performance data, as well as mean values. Decisions regarding apparent superiorities of algorithms or products cannot be reliably made if the differences in mean values are smaller than the associated variances. Reference to handbooks of experimental statistics can be valuable in avoiding misinterpretation of the test data.

In principle, extensive and statistically valid testing involves the use of large data bases. However, the costs of testing and resources required for these tests are frequently regarded as prohibitive, and more limited testing is typical. Consequently, attempts should be made to determine the statistical validity of the tests as an important factor in performance assessment.

Documentation

Proper documentation is an essential component in performance assessment. As recommended in this paper, information should be provided to document the relevant characteristics of the test speaker population, test vocabulary and other test data to establish the relevant context of the testing.

Test material obtained in accurate simulations of field applications may contain noise or speaker artifacts such as coughs, stammers, or false starts. Alternatively these artifacts may have been manually deleted or edited from the test and/or training material. The process of selection or preparation of the test material should be described.

Summary test data that should be documented include those in table 2.

These results are the most frequently cited performance data, however, documentation of confusion pairs (e.g., "five" and "nine") or confusion matrices is informative and provides useful information in designing applications vocabulary.

Response time is a critical factor in successful implementations of large vocabulary systems. There is no

Table 2. Data requiring documentation.

Correct Recognition Percent (Recognition Accuracy)	$= \frac{(\#\textit{Correctly Recognized Words})\times 100}{(\#\text{Test Words})}$ (Percent)
Substitution Percent	$= \frac{(\#\textit{Substituted Words})\times 100}{(\#\text{Test Words})}$ (Percent)
Deletion Percent	$= \frac{(\#\textit{Deleted Words})\times 100}{(\#\text{Test Words})}$ (Percent)
Insertion Percent	$= \frac{(\#\textit{Inserted Words})\times 100}{(\#\text{Test Words})}$ (Percent)
If the reject capability is employed, the following data are important:	
Rejection Percent	$= \frac{(\#\textit{Rejection Responses})\times 100}{(\#\text{Test Words})}$ (Percent)
Ratio of Total Errors to Rejections	$= \frac{(\#\textit{Substitutions} + \#\textit{Deletions} + \#\textit{Insertions})}{(\#\text{Rejections})}$
Settings of the Reject Threshold or Reject Criteria:	

accepted procedure for precise measurement and specification of response time for connected word systems. If the processing time is comparable to or less than the utterance duration, the system response time may be described as "real time." A suggested comparative measure for other than real time systems is a multiple of utterance duration, assuming processing is initiated at the beginning of the utterance and completed with display or return of the recognized word.

In view of the fact that processing times are finite, errors which arise from utterances spoken with insufficient pauses between words—(for isolated word systems, in particular) should be identified and noted in the performance documentation.

The A-weighted sound level (dBA) measured in the vicinity of the test speaker should be specified if environmental noise is believed to be a significant limitation on performance. More thorough documentation of the properties of the environmental noise may be appropriate.

Suggested documentation of the speech signal-to-noise properties of the test material should include the ratio of speech peak level to steady background noise level (in dB) measured according to ANSI S.3.59 [5]. If this is not feasible, at least, the range of typical maximum speech level to background level indication on the VU indicator of a conventional audio tape recorder should be noted and cited.

5. Perspectives on Testing

As explained in the preceding material, there is no simple and completely objective way to test the performance of automatic speech recognition technology. The number and complexity of factors influencing performance is such that in many cases, the relative advantages offered by competing algorithms, commercial products, or integrated systems may be obscured. The approach toward performance assessment in this paper emphasizes the value of benchmark tests and the need to carefully model applications. This is particularly important if the expenses of integrating speech technology into a particular, well-defined application and the benefits to be achieved are appreciable. Attention to detail in planning an applications test should be reflected in greater confidence in the ability of the technology to provide the anticipated benefits. Here also, a poorly-structured applications test or one that does not adequately account for important factors influencing performance will invalidate the test results and may lead to costly and unsuccessful attempts to use this new technology.

There are however, a number of other perspectives toward testing. It is useful to identify some of these perspectives.

Informal Device Testing

Within the past several years, the emergence of low cost commercial products has made an informal approach toward performance testing fairly widespread. Many individual purchasers of speech recognizers undertake an informal test program that primarily consists of familiarization of that purchaser or a designated individual with the technology. In many cases, the primary value of these tests appear to be that the experimenter learns to recognize the constraints imposed by the particular product on system enrollment, environmental factors, user interaction, interface design, etc. If informal testing of this sort results in development of a successful application, the experimenter gains valuable experience in the use of a new technology, insights into the selection of improved second-generation or competitive products and the design of more formal and reliable tests. However, there are real risks that the informal testing may not lead to the development of a successful application, that the purchaser may inappropriately conclude that the technology offers no promise for his application, and that the individuals involved in the testing and systems integration learn little of value in the process. Serious attention should be given to allocating adequate resources to carry out more formal tests to the point at which a serious and detailed investigation has taken place, and at which time the experimenter can demonstrate that he or she has developed an in-depth understanding of the relevant strengths and weaknesses. The information contained in this paper should be valuable in understanding the scope of the issues that should be addressed in these tests.

Workstation or Task Redesign

Another perspective toward studying the performance of this technology is based upon the desire to achieve the productivity benefits that might be offered by redesign of workstations or tasks, by using speech as an alternative or additional data entry or command/control modality. This is perhaps the implicit goal of all efforts to create successful applications. The design of tests to measure productivity benefits is beyond the scope of this report, but extremely important. In tests to measure these productivity benefits, those benefits which are specifically due to the use of speech technology should be compared to those that might be primarily due to redesign of workstations or tasks.

Human Factors Research

The design of successful human-machine dialogs is an active and important research topic at present. Not enough is known at present about the desired properties of automatic speech recognizers and the human-machine interaction to always lead to the design of successful applications. Further research on such issues as the design of optimal error correction protocols and user training and feedback is required and will serve to advance the technology. Attention to the factors contained in this report should serve to increase the value of these studies.

Common Concerns

The differing perspectives are consistent with the different outlooks regarding the purpose of the tests. The testing, like the technology itself, may serve different purposes. Whatever approach is taken, however, there are many common concerns including:

- Recognized complicating factors must be accounted for and carefully controlled. Failure to do so will invalidate the test data.
- Detailed documentation must be made available to indicate experimental design and to provide data and sufficiently detailed data analysis to indicate the significance of the test results to others. Detailed documentation should be part of reporting all performance tests. Failure to do so often leads to meaningless comparisons of product or system performance or misleading citations.
- Benchmark test data, in which severe constraints have been imposed on the test conditions, are extremely valuable. Typical constraints limit the test vocabulary, number of test talkers, format of the input speech, nature of environmental effects, and prohibit feedback between the user and the system. While it can be argued that under these conditions, it is possible to adapt or specially "tune" algorithm or device characteristics to optimize test performance for a specified test data base, benchmark tests provide data that are useful for initial comparisons of algorithm or device performance. Application testing should reveal whether or not the particular device selected for this phase of testing is well suited to a particular application or user's needs.

Agreement should be reached at an early stage of interest in the technology on the purpose of testing and appropriate measures of performance. Once these issues are decided, the nature of the tests can be detemined. In the absence of agreement on these issues, little progress can be made.

An *ad hoc* group met at the National Bureau of Standards in June 1982, to discuss performance assessment of speech recognition systems, following discussions during the NBS/NADC sponsored Workshop on Standardization for Speech I/O Technology. Further discussions were held at the 1982 (Paris), 1983 (Boston) and the 1984 (San Diego) meetings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP).

Following the 1982 ICASSP, the *ad hoc* group was constituted as the Speech I/O Technology Performance Evaluation Working Group, sponsored by the Speech Processing Technical Committee of the IEEE Acoustics, Speech and Signal Processing Society. Material presented in Section IV of this paper is adapted from informal drafts circulated within this Working Group.

Particular appreciation is expressed to Dr. Janet M. Baker, Chairman of the IEEE Speech I/O Technology Performance Evaluation Working Group, for her enthusiastic support and constructive criticism of this material, as well as to many other individuals who have shared their perspectives and expertise in addressing these issues.

6. References

[1] Reddy, R., and V. Zue, "Recognizing Continuous Speech Remains an Elusive Goal," *IEEE Spectrum*, November 1983.

[2] "Automatic Speech Recognition in Severe Environments," Report of the Committee on Computerized peech Recognition Technologies, National Research Council, National Academy Press (Washington, DC), 1984.

[3] Baker, J. M.; D. S. Pallett and J. S. Bridle, Speech Recognition Performance Assessments and Available Data Bases, *Proceedings of ICASSP83*, Boston, MA, April 14-16, 1983, p. 527-530.

[4] See for example, Burrington, R. S., and D. C. May, Handbook of Statistics with Tables (Second Edition) McGraw-Hill, New York, 1970, Sections 14.53-14.64.

[5] American National Standard Method of Measurement of Speaking Levels, ANSI S3.59 (pending).

7. Annotated Bibliography

The papers cited in this annotated bibliography include many of the key papers used for background material in preparing this paper. While there is very limited literature on the topic of performance assessment *per se*, there are valuable discussions of the several issues affecting performance in many of these papers. The interested reader is encouraged to refer to these key papers and the related literature cited therein in order to develop a more thorough understanding of the process of performance assessment.

Automatic-Speech and Speaker Recognition, N. R. Dixon and T. B. Martin, eds., IEEE Press (New York), 1979, p. 33. An important collection of papers that documents the state-of-the-art just prior to widespread use of VLSI technology in automatic speech recognition.

Baker, J. M., "The Performaing Arts—How to Measure Up!" Proceedings of the Workshop on Standardization for Speech I/O Technology, D. S. Pallett, ed., National Bureau of Standards, Gaithersburg, MD, March 18–19, 1982, p. 27–33. Presents the results of a series of tests on isolated word, speaker dependent recognizers, and a discussion of statistical factors in testing.

Baker, J. M., D. S. Pallett, and J. S. Bridle, "Speech Recognition Performance Assessment and Available Data Bases," Proceedings of ICASSP 83, Boston, MA, April 14–16, 1983, p. 527–530. Discusses data bases used in some well-documented benchmark tests.

Chollett, G. F. and C. Gagnoulet, "On the Evaluation of Speech Recognizers and Data Bases Using a Reference System," Proceedings of ICASSP 82, Paris, May 3–5, 1982, p. 2026–2029. Proposes the use of an openly published reference recognition algorithm to provide benchmarks of recognizer performance as well as of the relative difficulty of different speech vocabularies and data bases. Chollett credits R. K. Moore with early advocacy of this approach.

Chollett, G. F., and M. Rossi, "Evaluating the Performance of Speech Recognizers and Data Bases at the Acoustic-Phonetic Level," Proceedings of ICASSP 81, Atlanta, GA, March 30–April 1, 1981, p. 758–761. Advocates comparison of measured confusibility for recognizers with that of humans.

Clark, J., P. Collins, and B. Lowerre, "A formalization of performance specifications for discrete utterance recognition systems," Proceedings of ICASSP 81, Atlanta, GA, March 30–April 1, 1981, p. 753–757. Discusses needs for objective procedures for comparing the performance of isolated word recognizers.

DeMori, R., *Computer Models of Speech Using Fuzzy Algorithms*, Plenum Press (New York), 1983, p. 382–386. Proposes an experimental model for speech recognition and understanding using fuzzy-set theory and concepts of artificial intelligence. Includes a review of several procedures for evaluating language complexity.

Doddington, G. R., and T. B. Schalk, "Speech Recognition: Turning Theory into Practice," IEEE Spectrum, Sept. 1981., p. 26. Describes the results of a series of benchmark tests on isolated word, speaker dependent recognizers, and presents a thoughtful discussion of factors complicating the performance assessment test procedure.

Goodman, R. G., Analysis of Language for Man-Machine Communication, Ph.D. Dissertation, Stanford University, May 1976. An early study of the complexity of imposed grammars and the issues of phonetic ambiguity and measures of the similarity of acoustic events.

Jelinek, F., Mercer, R. L., Bahl, L. R., and Baker, J. K., "Perplexity—a measure of the difficulty of speech recognition tasks," J. Acoust. Soc. Am. Vol. 62, Suppl. No. 1, Fall 1977, p. S63. The abstract of a paper given at the Fall 1977 meeting of the Acoustical Society of America. This paper presented material summarized by Jelinek, Mercer, Bahl, below.

Jelinek, F., Mercer, R. L., Bahl, L. R., "Continuous Speech Recognition: Statistical Methods," Chapter 25 in Handbook of Statistics, Vol 2, P. R., Krishnaiah and L. N. Kanal, eds., North-Holland Publishing Co. (1982) 549–573. These papers describes methods of assessing the performance of continuous speech recognition systems and the relative difficulty of recognition tasks.

Lea, W. A., "What Causes Speech Recognizers to Make Mistakes?", Proceeding of ICASSP 82, Paris, May 3–5, 1982, p. 2030–2033. Discusses factors that affect recognition accuracy and cites other papers by this author on related topics.

Leonard, R. G., "A Database for Speaker-Independent Digit Recognition," Proceedings of ICASSP 84, San Diego, CA, March 19–21, 1984, p. 42.11.1–42.11.4. Describes valuable data base for research and test purposes that represents more than 20 dialects using approximately 100 adult males, 100 adult females and 100 children. Includes a discussion of an experimental procedure to "certify" the data base using human subjects.

Moore, R. K., "Evaluating Speech Recognizers," IEEE Transactions ASSP Vol 25, No. 2, 1977. A significant early contribution proposing a standard for comparing the performance of different recognizers based on a model of human word recognition.

Peckhan, J. B., "Speech Technology Assessment Activities in the U. K.," Proceedings of Speech Tech '85, New York, NY, April 22–24, 1985, pp. 165–169. Describes the activities and goals of the Speech Technology Assessment Group formed in the U. K. within the Speech Group of the Institute for Acoustics.

Poock, G. K., Martin, B. J. and Roland, E. F., The Effect of Feedback to Users of Voice Recognition Equipment, Naval Postgraduate School Report NPS55-83-003, February 1983. One of a series of studies conducted by Poock and his colleagues addressed at measurement of the performance of integrated or interactive systems.

Proceedings of the Workshop on Standardization for Speech I/O Technbology, D. S. Pallett, ed., National Bureau of Standards, Gaithersburg, MD, March 18–19, 1982. Presents approximately thirty technical papers addressing the issue of performance assessment for speech I/O technology.

Proceeding of the Voice Data Entry Systems Applications Conference '84, L. Lerman, ed., American Voice Input/Output Society, Arlington, VA, Sept. 11–13, 1984. Most recent in a series of four Conferences addressing the design of applications of voice-interactive systems for data entry and other purposes.

Rosenberg, A. E., " A probabilistic model for the performance of word recognizers," AT&T Technical Journal, Vol. 63, No. 1, Jan. 1984, pp. 1–32. Develops analytic models of recognizer performance.

Sondhi, M. M., and Levinson, S. E., "Relative difficulty and robustness of speech recognition tasks that use grammatical constraints," J. Acoust. Soc. Am., Vol. 62, Suppl. No. 1, Fall 1977, p. S64. The abstract of a paper given at the Fall 1977 meeting of the Acoustical Society of America. This paper presented material summarized by Sondhi and Levinson, below.

Sondhi, M. M., and Levinson, S. E., "Computing Relative Redundancy to Measure Grammatical Constraint in Speech Recognition Tasks," Proceeding of ICASSP 78, Tulsa, OK, May 1978, pp. 409–412. This paper develops algorithms for computing various statistical properties of finite languages and documents these properties for eight speech recognition task languages that have appeared in the literature.

Spine, T. M., Williges, B. H. and Maynard, J. F., "An Economical Approach to Modeling Speech Recognition Accuracy," Int. J. Man-Machine Studies Vol 21, 1984, pp. 191–202. Discusses the use of a central-composite design methodology as a means to develop empirical prediction equations for speech recognizer performance. Factors manipulated in the experimental design include number of training passes, reject threshold, difference score and size of the active vocabulary.

Wilpon, J. G. and Rabiner, L. R., "On the Recognition of Isolated Digits from a Large Telephone Customer Population," Bell System Technical Journal, Vol. 62, No. 7, September 1983, pp. 1977-2000. Describes the collection and use of an isolated digit data base for use in research and development of speaker independent speech recognition systems.

Wilpon, J.G., "A Study on the Ability to Automatically Recognize Telephone Quality Speech from Large Customer Populations," AT&T Technical Journal, Vol. 64, No. 2, February 1985, pp. 423-451. Provides an important and detailed analysis of procedures used in collecting speech data bases over telephone lines. Together with the preceding paper by Wilpon and Rabiner, valuable information is presented on the complexity of successful implementation of speaker-independent automatic speech recognition over telephone lines.

Appendix: Terminology

Since automatic speech recognition is an emerging technology, a standard terminology has not yet been established. Current activities within the IEEE Acoustics, Speech and Signal Processing Society include a Working Group on Speech I/O Systems Performance Assessment. This Working Group has discussed the desirability of use of a uniform terminology in technical papers, presentations, and vendor's specifications, and have contributed to the suggested definitions of terminology contained in this Appendix and used in this paper.

Active Vocabulary—See "Vocabulary"

Adaptation—The automatic modification of existing internal machine representations (e.g., template sets, word models, etc.) of specific utterances and/or noise.

Artificial Language—See "Constrained Language"

Automatic Speech Recognition—The process or technology which accepts speech as input and determines what was spoken.

Automatic Speech Recognition System—An implementation of algorithms accepting speech as input and determining what was spoken.

Automatic Speech Recognizer—A device implementing algorithms for accepting speech as input, determining what was spoken, and providing potentially useful output depending on word(s) recognized.

Connected Words—Words spoken carefully, but with no explicit pauses between them.

Constrained Language—Lexically and syntactically constrained word sequences (e.g., telephone numbers).

Continuous Speech—Words spoken fluently and rapidly as in conversational speech.

Deletion—An instance in which a spoken word is ignored, and for which the recognizer or system provides no response (e.g., in recognizing a string of digits, if the recognizer returns one less digit than has been input).

Discrete Utterance Recognition—The process of recognizing a word or several words spoken as a single entry.

Enrollment—The process of constructing representations of speech, such as template sets or word models, to be used by a recognizer. Also referred to as "system training," as distinct from "user training."

Enrollment Data—See "Training Data"

False Acceptance—An example of failure to reject properly spoken input utterances that are not part of the active vocabulary, resulting in selection of a word in the active vocabulary.

Grammar—In general, a grammar of a language is a scheme for specifying the sentences allowed in the language, indicating the rules for combining words into phrases and clauses. In automatic speech recognition, task grammars specify the active vocabularies and the transition rules that define the sets of valid statements to complete the tasks. The task grammar and structured vocabulary provide syntactic control of the speech recognition process that can greatly enhance performance.

Insertion—An instance of a recognition occurring due to spurious noise or an utterance other than those that are legitimate on syntactic considerations. In the former case, some input other than an utterance (typically some ambient or electrical noise artifact) is not properly rejected and the system response indicates that some utterance in the recognition vocabulary occurred. In the latter case, a word that has been uttered (but which is not part of the active recognition vocabulary because of current syntactic constraints) is falsely accepted as an utterance from the active recognition vocabulary.

Isolated Words—Words spoken with pauses (typically with duration in excess of 200 ms) before and after each words.

Isolated Word—See "Discrete Utterance Recognition"

Natural Language—Syntactically unconstrained word sequences, typically drawn from a large lexicon and complying with conventional usage.

Practice Data—Any speech material (utterances) used in developing a recognition system prior to a test of that particular recognizer.

Recognition Systems—See "Automatic Speech Recognition Systems"

Recognition Unit—The basic unit of speech on which recognitions being performed, often presumed to be the word. The actual unit used may be smaller (e.g., phones, demisyllables, syllables or features) or larger (e.g., multi-word phrases or utterances).

Recognition Vocabulary—See "Vocabulary"

Recognizer—See "Automatic Speech Recognizer"

Rejection—The property of rejecting inputs. There are three general classes of system response involving rejection: i) noise rejection, ii) rejection of improperly spoken input utterances, iii) rejection of properly spoken input utterances that are part of the active vocabulary, sometimes termed false rejection.

Speaker Dependent Recognition—A procedure for speech recognition which depends on enrollment data from the individual speaker who is to use the device.

Speaker Independent Recognition—A procedure for speech recognition which requires no previous enrollment data from the individual speaker who is to use the device.

Speech Level—A logarithmically based measure of the amplitude of a speech waveform. Accurate specification of speech level is important in specifying the input signal amplitude when testing recognizers and when specifying signal-to-noise ratio. American National Standard ANSI S3.59 provides a well-specified procedure for measurement of speech level.

String—A sequence of spokenwords or phrases, often spoken as connected words or continuous speech and intended to provide a single useful input to a recognizer (e.g., a five-digit ZIP Code or a seven-digit telephone number).

Substitution—An instance in which one word in the recognition vocabulary is incorrectly recognized as another word in the recognition vocabulary.

Syntax—Structure by which grammatical word sequences are specified.

Test Data—Any speech material (utterances) used in a particular test of a recognizer not previously used in developing or modifying that recognizer. The same set of test data may be used repeatedly for tests of different recognizers or in production testing, but not for continuing tests of an algorithm or recognizer in development.

Token—A sample speech utterance.

Training—See "Enrollment." "System training" is preferably referred to as "enrollment." "User training" refers to the process of user familiarization with speech technology (e.g., learning how to use an automatic speech recognition device).

Training Data—Speech material used to construct parametric representations of speech such as template sets or word models used by a recognizer. Also referred to as enrollment data. Not to be confused with performance data obtained in training potential users of the technology.

Utterance—A word or multi-word phrase spoken continuously as a single unit.

Vocabulary—The words or phrases to be recognized by a recognizer. Distinctions should be made between the complete set of all words or phrases that a recognizer has been trained or programmed to recognize, sometimes called the total *recognition vocabulary*, and the (instantaneously varying) subset of these that may be active at a given time because of an imposed task grammar or other syntactic constraint, called the *active vocabulary*.

Word—See "Recognition Unit"

Word Model—A parametric (coded) representation of the sound patterns of words as a sequence of units such as phonetic units, syllables, or other speech parameters.