ELSEVIER

Speech Communication 22 (1997) 1–15

SPEECH COMMUNICATION

Speech recognition by machines and humans

Richard P. Lippmann [*]

Lincoln Laboratory MIT, Room S4-121, 244 Wood Street, Lexington, MA 02173-9108, USA

Received 2 February 1996; revised 14 November 1996; accepted 28 April 1997

Abstract

This paper reviews past work comparing modern speech recognition systems and humans to determine how far recent dramatic advances in technology have progressed towards the goal of human-like performance. Comparisons use six modern speech corpora with vocabularies ranging from 10 to more than 65,000 words and content ranging from read isolated words to spontaneous conversations. Error rates of machines are often more than an order of magnitude greater than those of humans for quiet, wideband, read speech. Machine performance degrades further below that of humans in noise, with channel variability, and for spontaneous speech. Humans can also recognize quiet, clearly spoken nonsense syllables and nonsense sentences with little high-level grammatical information. These comparisons suggest that the human–machine performance gap can be reduced by basic research on improving low-level acoustic-phonetic modeling, on improving robustness with noise and channel variability, and on more accurately modeling spontaneous speech. © 1997 Elsevier Science B.V.

Résumé

Ce papier présente un bilan des travaux comparant les performances des systèmes de reconnaissance de parole modernes à celles des locuteurs humains. Les comparaisons sont basées sur six types de corpus de parole avec des vocabulaires allant de 10 à plus de 65 000 mots et des contenus allant des mots isolés à des conversations spontanées. Les taux d'erreurs des machines sont souvent supérieures de plus d'un ordre de grandeur à celles des humains pour la parole lue en atmosphère calme et transmise en large-bande. Les performances des machines se dégradent encore par rapport à celles des humains dans les contextes bruités, ou de qualité de transmission variable et pour la parole spontanée. Les locuteurs humains peuvent également reconnaitre, avec peu d'information linguistique de haut-niveau, des syllabes ou des phrases sans signification quand elles sont prononcées clairement dans des atmosphères calmes. Ces comparaisons suggèrent que l'écart important qui subsiste entre les performances des machines et celles des humains peut être réduit par des recherches de base sur les sujets suivants: l'amélioration de la modélisation acoustico-phonétique de bas-niveau, l'amélioration de la robustesse au bruit et à la variabilité des conditions de transmission, et la modélisation plus précise de la parole spontanée. © 1997 Elsevier Science B.V.

Keywords: Speech recognition; Speech perception; Speech; Perception; Automatic speech recognition; Machine recognition; Performance; Noise; Nonsense syllables; Nonsense sentences

[*] E-email: rpl@sst.ll.mit.edu.

PII S0167-6393(97)00021-6

1. Introduction

Dramatic advances have been made in speech recognition technology over the past few years. Vocabulary sizes now exceed 65,000 words and fast decoding algorithms allow continuous-speech recognition systems to provide near real-time response. Despite these advances, commercial recognizers have been successful only in a few constrained application areas. Many researchers believe that recognizers will enjoy widespread use and become commonplace only if their performance approaches that of humans under everyday listening environments. This paper measures how far research has progressed towards this goal. Results from scattered studies which have compared human and machine speech recognition on similar tasks are summarized to determine how much speech recognizers must improve to match human performance. Speech corpora used in these comparisons do not represent everyday listening conditions, but they span a continuum ranging from quiet read isolated words, to noisy read sentences, to spontaneous telephone speech. Results clearly demonstrate that modern speech recognizers still perform much worse than humans, both with wideband speech read in quiet, and with band-limited or noisy spontaneous speech.

Results comparing humans to machines are presented with four important goals. These are to motivate research in directions that will decrease the human–machine performance gap, to promote further human–machine comparisons, to promote further experimental work with human listeners to understand how humans adapt to talker and environmental variability, and to encourage a multi-disciplinary dialog between machine recognition and speech perception researchers. In the remainder of this paper, Section 2 describes six modern speech corpora used to evaluate machine recognizers, Section 3 discusses some important issues involved in comparing human and machine performance, Section 4 presents human and machine error rates for the six corpora, and Section 5 presents a summary and discussion.

2. Talker-independent speech recognition corpora

The six speech corpora shown in Fig. 1 were created to develop and evaluate machine speech recognizers. Human and machine performance can be compared using the many machine results obtained using these corpora and human recognition studies obtained with these or similar speech materials. These corpora span a wide range of difficulty and represent many different potential applications of speech recognition technology. All are designed to test talker-independent recognition of speech from talkers not used for training. Materials were recorded by prompting talkers to produce words in isolation, by having talkers read carefully prepared sentences, and also by recording extemporaneous telephone conversations on various topics including ‘‘credit cards’’. Vocabulary sizes range from 10 to 5,000 words for the smaller corpora. Recent tests using the North American Business News (NAB) and the Switchboard corpus use an unlimited vocabulary size and often employ machine recognizers with vocabularies that exceed 65,000 words (e.g. Woodland et al., 1996). In these tests, out-of-vocabulary errors are introduced by the small numbers of test words that are not contained in the vocabularies of machine recognizers.

Most corpora focus on a dictation task where the goal is to identify all spoken words. Word error rates reported for these corpora treat substitutions, deletions, and word insertions as errors. The Switchboard corpus, however, was used both for dictation and also to determine the ability of machine wordspotters to detect 20 common words in conversational telephone speech. Wordspotters are used for computer and telephone interfaces with untrained users because they do not rely on strong grammars to pro-

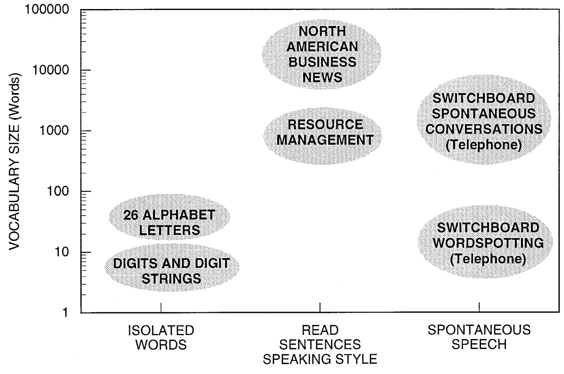

Fig. 1. Six speech recognition corpora used to compare humans and machines have vocabularies ranging from 10 to more than 65,000 words and contain both spontaneous conversations and read isolated words or sentences.

vide good performance. The performance metric used for wordspotting is the average detection rate, or the percentage of true keyword occurrences that are detected. The ''miss'' rate is 100 minus the detection rate. The detection rate is averaged over wordspotter systems adjusted to provide false alarm rates ranging from 1 to 10 false alarms per keyword per hour of conversational speech.

Characteristics of the six speech corpora are provided in Table 1. Data in Table 1 refer either to the total amount of spoken speech data contained in each corpus (TI Digits, Alphabet Letters, Switchboard) or to the amount of speech data available for training (all other corpora). Ranges are provided in columns three through five for the NAB and Switchboard corpora because these corpora have grown in size over the years. Early smaller versions were used for initial tests, and larger versions are being used for more recent evaluations. The column labeled ''perplexity'' is provided because perplexity is a much better predictor of recognition difficulty than vocabulary size. This statistical measure can be thought of as the average number of alternative words which the recognizer must choose between or the ''branching factor'' of the language model used by machine recognizers (e.g. Jelinek, 1985). For all machine recognizers, perplexity is determined both by the language model and the materials used for testing. The value of perplexity is 10 for digit recognizers which treat each digit as equally likely, independent of the surrounding digits. Perplexity increases to values ranging from 45 to 160 for read sentences from business publications and spontaneous speech. The maximum perplexity of 1,000 is obtained by recognizers for the 1,000-word Resource Management task when each word is treated as equally likely. High-performance continuous-speech recognizers for the NAB and Switchboard corpora often use trigram language models which estimate probabilities of the next word in a sequence using the identity of only the two previous words (e.g. Jelinek, 1985). Free parameters in trigram grammar models can be estimated using textual materials containing more than 300 million words of text that are now provided along with the speech materials in the NAB corpus (Stern, 1996).

These corpora include a wide range of speech materials. The TI-digits corpus (Leonard, 1984) contains isolated digits and two-to-seven digit sequences. The alphabet letters corpus (Cole et al., 1990) contains isolated spoken letters which could be used to spell a person's name. The Resource Management corpus (Price et al., 1988) contains highly constrained sentences that can be used to query a naval data base. Two sample sentences are ''*Are there fourteen ships at sea*'' and ''*List all cruisers and their fleet identifications*''. The NAB corpus originally contained only sentences selected from articles in the *Wall Street Journal* (Paul and Baker, 1992). In recent years it has been expanded to contain sentences from other business publications (Kubala, 1995; Stern, 1996). Two sample sentences are ''*For the first time in years the Republicans also captured both houses of Congress*'' and ''*Terms weren't disclosed, but industry sources said the price was about $2.5 million*''.

The NIST Switchboard corpus (Godfrey et al., 1992; LDC, 1995) has been used both for wordspot-

Table 1
Characteristics of six talker-independent speech recognition corpora

Corpus	Description	Numbers of talkers	Vocabulary size	Number of utterances	Total duration	Recognition perplexity
TI digits	Read digits	326	10	25,102	4 hrs	10
Alphabet letters	Read alphabet letters	150	26	7,800	1 hr	26
Resource Management	Read sentences	109	1,000	4,000	4 hrs	60–1,000
North American Business News (NAB)	Read sentences	84–284	5,000 – Unlimited	7,200–37,200	12–62 hrs	45–160
Switchboard continuous speech recognition	Spontaneous telephone conversations	70–543	2,000 – Unlimited	35–2400 conversations	2–240 hrs	80–150
Switchboard wordspotting	Spontaneous telephone conversations	70	20 keywords	2,000 keyword occurrences	2 hrs	–

ting and continuous speech recognition. It contains speech recorded over normal telephone lines from two talkers located at home or work who were instructed to carry on a spontaneous conversation concerning a specific topic. One component of this corpus called the Credit Card corpus contains conversations concerning the topic of credit cards. Two sample excerpts are ‘‘*I don't know if I'm really afraid of spending too much*’’ and ‘‘*I uh, I try to get maybe just one or two*’’. These phrases contain false starts and are frequently non-grammatical due to the spontaneous nature of the conversations. This material also samples a wide range of talking styles and acoustic environments. Wordspotter tests with the Credit Card corpus search for 20 frequently occurring keywords including ‘‘card’’, ‘‘credit’’ and ‘‘charge’’. Continuous speech recognition tests recognize all spoken words.

3. Issues in comparing the performance of machines and humans

The performance of machines and humans is compared in this paper using only word error rates because other measures such as training time, recognition time, and amount of prior information required concerning the task such as vocabulary size, type of additive noise, bandwidth, and semantic context is difficult to measure and compare. The lowest machine and human error rates are always reported. This involves finding the lowest machine error rates from the recent literature for each speech corpus and also finding the lowest human error rates reported for each corpus or for a similar task. In most experiments, human listening tests were performed with materials from the same speech corpus used for testing machine recognizers. Only for the Resource Management corpus with a null grammar and for the Alphabet Letters corpus, are machine results compared to human results obtained with other materials. These comparisons use materials with roughly the same perplexity and acoustic characteristics.

Human error rates would ideally be compared only to error rates of the current single best-performing machine recognizer. Such comparisons are impossible because no single machine recognizer has been tested on all corpora and no single recognizer has been shown to provide best performance across all conditions sampled by these corpora. Comparisons provided in this paper inflate machine performance because they use the machine recognizer which provided the best performance on each corpus, and no single machine recognizer can uniformly obtain such low scores. In general, recognizers developed for one corpus degrade on other corpora due to the careful tuning required to provide best performance. Tuning includes limiting the recognition vocabulary, testing using a constrained grammar, selecting input features, adjusting free grammar parameters, adjusting noise and channel adaptation algorithms, and training under conditions similar to those expected during testing. Tuning may lead to only small differences in error rates across corpora when speech corpora are similar (e.g. *Wall Street Journal* sentences read by Native English versus Native American talkers). It can also lead to large variations in error rate when corpora are more dissimilar (e.g. read wideband *Wall Street Journal* sentences versus spontaneous Switchboard telephone conversations).

Results obtained with human listeners are also sometimes inflated because experimenters reduced errors caused by inattention by allowing multiple listening passes and by using a committee majority vote across multiple listeners instead of using the average individual error rate. Experimenters also sometimes corrected for spelling errors to eliminate out-of-vocabulary responses. Such out-of-vocabulary responses are not possible for machine recognizers which use limited-vocabulary testing when the training and testing vocabularies are known and identical. Spelling corrections typically correct proper nouns such as ‘‘Fannie Mae’’ and ‘‘Shearson Lehman’’ and sometimes halve human error rates (Ebel and Picone, 1995). Using a committee majority vote to eliminate errors caused by inattention also sometimes halves the human error rate (e.g. Ebel and Picone, 1995).

In all experiments described in this paper, human listeners always first verify the accuracy of speech materials by creating a text transcription before these materials are then used with other listeners for intelligibility testing. Valid intelligibility testing requires accurate transcription. Transcription accuracy is maintained at a high level high first because listeners used during transcription are generally highly-motivated (they are often the experimenters), because they typically only make a simple binary decision and decide whether a talker produced the text used for prompting, and because any questionable speech tokens can be eliminated from intelligibility test experiments. The best listening conditions are also used during transcription. Speech is presented before artificially adding noise or other channel degradations and, in noisy or reverberant environments, an additional reference microphone positioned close to the talker is often provided to obtain high-quality speech for transcription (e.g. Pallett et al., 1995). In addition, a committee of motivated listeners is often used to transcribe some of the more difficult materials (e.g. Leonard, 1984; Martin, 1996). If it is assumed that the transcription word error rate is lower than reported word error rates for intelligibility testing with human listeners, then the transcription error rate is less than 0.009% for read digits (Leonard, 1984), less than 0.4% for read sentences from the *Wall Street Journal* (Ebel and Picone, 1995), and less than 4% for spontaneous conversations recorded over the telephone (Martin, 1996).

4. Performance comparisons

Sections 4.1, 4.2, 4.3, 4.4 and 4.5 summarize recent comparisons between the speech recognition performance of humans and machines using the six speech corpora shown in Fig. 1 supplemented by other speech materials that have been used to evaluate human performance. Studies using digit strings and isolated letters are summarized in Section 4.1. Reviews of studies that evaluated continuous-speech recognition performance with the Resource Management and NAB read-sentence corpora and with the Switchboard spontaneous speech corpus are presented in Sections 4.2, 4.3 and 4.4. Section 4.5 summarizes studies which used the Switchboard spontaneous speech corpus for wordspotting.

4.1. Digits and alphabet letters

The Texas Instruments digit corpus (Leonard, 1984) which contains more than 25,000 utterances has been used extensively to evaluate the performance of talker-independent machine speech recognizers designed to recognize isolated digits and digit strings. Human error rates were obtained using 26 highly motivated listeners who initially listened to degraded 12th-order LPC synthesized versions of the original data, received bonus pay when they made fewer errors, entered responses on a keyboard, and could listen to segments repeatedly (Leonard, 1984). The average digit-string or per-utterance error rate for listeners presented with vocoded speech was 0.105%. The string error rate for vocoded speech dropped to 0.01% when the majority vote from a committee of three listeners was used to measure performance. The intelligibility of original unvocoded speech was estimated by presenting three listeners with wideband speech at a sample rate of 20 kHz. They only listened to those 70 tokens which were misclassified at least once when vocoded. The average string error rate for individual listeners was reduced to 0.009% using wideband speech. This error rate is plotted in Fig. 2 along with the average human string error rate for vocoded speech and the machine recognition string error rate from (Chou et al., 1994) obtained using a hidden Markov model (HMM) recognizer designed specifically for this task. This machine recognizer has the lowest published error rate on this corpus. Fig. 2 demonstrates that highly motivated human performance on this task is extremely good. The human error rate of 0.009% for wideband speech represents an average of 2 to 3 errors per listener over more than 25,000 tokens. Machine performance at 0.72% is almost two orders of magnitude worse. Human performance degrades with vocoded speech, but human error rates still remain roughly a factor of 7 lower than machine error rates.

The human error rate for continuously spoken letters of the alphabet and the machine error rate for

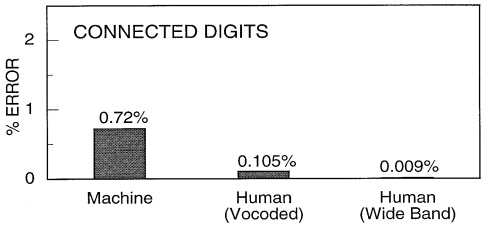

Fig. 2. Human and machine error rates for the digit corpus (Chou et al., 1994; Leonard, 1984).

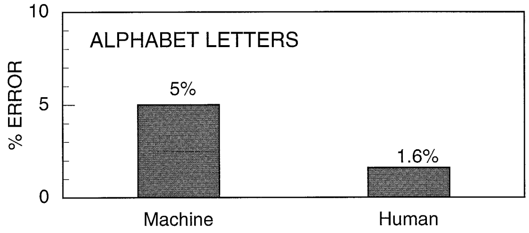

Fig. 3. Human error rates for continuously spoken letters (Daly, 1987) and machine error rates for isolated spoken letters (Cole et al., 1990).

isolated letters from the Alphabet Letters corpus described in Table 1 are presented in Fig. 3. The 1.6% human error rate shown in Fig. 3 is an upper bound for isolated letters because it was measured using more difficult continuously spoken letters. It was obtained using a corpus of spelled English words and nonsense words with 8 listeners and materials produced by 10 male and 10 female talkers (Daly, 1987). Talkers produced a total of 1,000 sequences (350 words and 650 nonsense words) of 3–8 continuously spoken letters by spelling letters in each word. The average error rate of letters for listeners was 1.6% both for meaningful words and nonsense words and the standard deviation across listeners was low (0.72 percentage points). This low error rate for continuously produced sequences of letters is similar to the low human error rates that have been reported for isolated nonsense syllables presented to well-trained and motivated listeners (Fletcher and Steinberg, 1929; Lippmann et al., 1981). Early intelligibility testing experiments at Bell Laboratories used consonant–vowel–consonant (CVC) nonsense syllables and crews of typically eight well-trained subjects serving as both listeners and talkers. These talkers performed live tests using lists of 66 CVCs each constructed by randomly choosing from 22 consonants and 11 vowels. The average syllable error rate in a quiet environment with direct air transmission was 2.0% with a standard deviation of 0.4 percentage points (Fletcher and Steinberg, 1929). The consonant error rate measured in similar experiments for consonant–vowel (CV) and vowel–consonant (VC) nonsense syllables constructed from 25 consonants and 3 vowels was only 0.5% (Fletcher and Steinberg, 1929). More recent experiments used CVC nonsense syllables formed by randomly choosing from 16 consonants and 6 vowels. The average syllable error rate for three listeners and 750 syllables with quiet wideband speech was only 1.5% (Lippmann et al., 1981).

The isolated-letter machine error rate in Fig. 3 for the Alphabet Letters corpus is from a neural network recognizer designed specifically for recognizing isolated letters (Cole et al., 1990). This recognizer was trained using one repetition of each letter from 60 male and 60 female talkers (3,120 tokens) and then tested using two repetitions of each letter from 30 different talkers (1,560 tokens). It has the best published performance on this task. As can be seen, the 1.6% human error rate for continuously spoken letters is roughly three times lower than the 5% machine recognition error rate for isolated letters.

The spoken letter corpus used to obtain the human error rate in Fig. 3 was also used for spectrogram reading experiments (Daly, 1987). Six trained spectrogram readers transcribed 100 utterances each (five from each of the 20 talkers) containing roughly 1/3 real words and 2/3 nonsense words. They were told that some utterances were real words, but were not told the exact proportion of real to nonsense words. They were also provided the talker identity for each utterance and other gross characteristics of the spoken letter corpus. Spectrogram readers provided a letter sequence for each spectrogram to permit comparisons to human listening experiments. The average error rate across all 5,601 letters was 9% ranging from 5% to 14% across spectrogram readers. There was also a wide range in the error rate across the 20 talkers (2.4% to 16%), female utterances were more difficult than those produced by males, and real words were generally easier to transcribe than nonwords. These results demonstrate that sequences of spoken letters can be recognized using only low-level acoustic-phonetic information conveyed to humans using a spectrographic display. The 5% error rate for the best spectrogram reader on this task is good, but still substantially worse than the 1.6% error rate for the average human listener.

4.2. Resource management

The Resource Management corpus was the first large-vocabulary corpus that was widely used for talker-independent continuous speech recognition. Over a period of roughly four years, machine error rates on this corpus using a word-pair grammar with a perplexity of 60 dropped from roughly 20% (Lee, 1989) to 3.6% (Huang et al., 1991; Pallett, 1991). One study measured the human word recognition error rate for a single listener using the 300 1989 Resource Management test sentences (Zue et al., 1989). This error rate was 0.1% corresponding to missing only 3 out of 2561 words. This error rate was not obtained using many listeners, but it is substantially lower than the best reported machine error rate of 3.6%.

Machine error rates increase dramatically for the Resource Management task when a ‘‘null grammar’’ is used where a machine recognizer assigns equal probability to all words independent of the surrounding words. This results in a much more difficult recognition task with a test perplexity bounded above by 1,000. The effect of using a null grammar in a speech recognizer instead of a stronger grammar is roughly equivalent to the effect of using nonsense sentences instead of meaningful English sentences with human listeners. The null grammar recognition condition uses a recognizer which makes little use of word sequence information, while nonsense sentences provide human listeners with limited word sequence information.

Nonsense sentences are semantically meaningless word sequences created by randomly selecting keywords and placing them in slots within a sentence frame. Many different types of nonsense sentences have been created for speech perception research (e.g. Lippmann et al., 1981; Miller, 1962). One set of nonsense sentence materials which uses a vocabulary size that is similar to that of the Resource Management task was created by randomly selecting words from 1,000 phonetically balanced monosyllabic words and inserting them in the sentence frame ‘‘The ____ ____ told the ____ ____’’, as described in (Lippmann et al., 1981). An example sentence is ‘‘*The cuff golf* told the *hold dive*’’. An average word error rate of 2.0% was obtained for the inserted words using 600 sentences produced by one male and one female talker and presented to three listeners. These listeners were provided little training, knew that words in sentences were selected at random, used an unrestricted response vocabulary, listened to each sentence twice, and were allowed more than 10 seconds for written responses. Words in these sentences are produced at a faster rate than in isolation and with cross-word coarticulation, but with limited contextual information of the type that is contained in the grammar models used in machine recognizers. Humans can still, however, make use of lexical information and of phonotactic constraints which characterize normal English speech to aid recognition. An older study also used nonsense sentences to study the importance of context for speech perception (Miller, 1962). Both nonsense and meaningful sentences were created using a small, 25-word vocabulary. Nonsense sentences such as ‘‘Socks wrong the has he’’ were creating by reversing the word order of meaningful sentences such as ‘‘He has the wrong socks’’. The intelligibility of words in nonsense sentences was high (word error rate less than 2% at high SNRs) and similar to the intelligibility of words presented in isolation, as long as a 10 second pause was provided between successive sentences to allow sufficient time for an oral response.

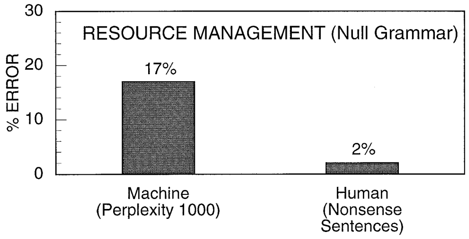

Fig. 4. Machine error rates for the Resource Management corpus with a null grammar and a vocabulary size of 1,000 (Huang et al., 1991; Pallett, 1991) and human error rates for nonsense sentences formed using a 1,000 word vocabulary (Lippmann et al., 1981).

Fig. 4 compares error rates obtained using the best performing null-grammar HMM recognizer on the Resource Management corpus (Huang et al., 1991; Pallett, 1991) to human error rates for nonsense sentences (Lippmann et al, 1981). Machine error rates are for the restricted 1,000-word Resource Management corpus using a highly tuned HMM recognizer. The 2% human word error rate is almost an order of magnitude lower than the 17% null grammar machine word error rate. These results demonstrate that humans can accurately perceive wideband speech in a quiet environment with limited contextual information using primarily low-level acoustic-phonetic information. The representative high-performance HMM recognizer described in (Huang et al., 1991), however, has much worse low-level acoustic modeling and relies heavily on contextual information and a constraining grammar to achieve good performance. Its error rate drops to 3.6% when recognition perplexity is reduced to 60 using word-pair grammar (Pallett, 1991). The relative contribution to recognition performance provided by a constraining grammar and lower-level modeling has not been assessed recently using null grammars because modern 5,000 to 65,000 word vocabulary continuous speech recognizers require a low-perplexity grammar to make machine recognition computationally tractable.

4.3. North American business news

The initial *Wall Street Journal* component of the NAB corpus provided researchers working on large-vocabulary continuous-speech recognition with a much more challenging task than the artificial Resource Management task. This corpus contains a larger vocabulary and uses read sentences prompted using text extracted from a published journal. Although many researchers have evaluated machine recognizers with this corpus, few studies have been performed with human listeners. A recent human listening study used *Wall Street Journal* sentences from one condition (Spoke 10) of the 1994 ARPA continuous speech recognition evaluation where performance was measured in quiet and at three speech-to-noise ratios (SNRs) using a 5,000 word vocabulary (Pallett et al., 1995; Kubala 1995). Noise was recorded in an automobile traveling at roughly 60 miles per hour and sentences were recorded with a high-quality close-talking microphone in a different quiet environment. Noise was artificially added to sentences, and segments containing only noise were provided for training HMM recognizers and developing new algorithms to adapt recognizers to the noise environments. Human listening experiments described in (Ebel and Picone, 1995) used 12 normal hearing college-educated adults who could listen repeatedly to any one of the 113 test sentences and entered responses using a common computer text editor with an unrestricted response vocabulary. Obvious typing errors, spelling errors, and out-of-vocabulary words were corrected. For example, the response ‘‘*Fanny May*’’ was converted to ‘‘*Fannie Mae*’’ and the name ‘‘*Sheerson Leeman*’’ was converted to ‘‘*Shearson Lehman*’’. A group of at least three listeners provided responses to the same sentences and a committee response was computed using a majority vote over these three listeners to eliminate errors caused by inattention. Although this procedure halved the error rate, the more conservative average error rate across listeners is reported here.

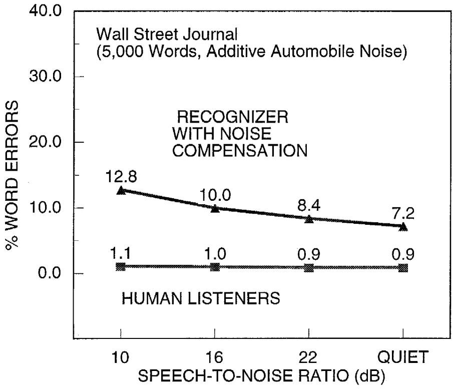

Fig. 5. Performance of humans (Ebel and Picone, 1995) and of a high-performance HMM recognizer with noise compensation (Gopinath et al., 1995; Pallett et al., 1995) for *Wall Street Journal* sentences with additive automobile noise.

Fig. 5 shows average Spoke 10 word error rates for humans and machines. The machine error rates are for an adaptation algorithm described in (Gopinath et al., 1995) which provides good performance and requires only a few seconds of noise-only adaptation data. This adaptation algorithm was essential for obtaining good machine performance. Machine error rates without adaptation exceeded 40% at SNRs of 10 and 16 dB. Adaptation requires two passes through each sentence. One pass is used to determine how HMM spectral features trained in quiet should be modified to make them similar to noisy spectral features. A second pass then performs recognition using the modified features. Human error rates from (Ebel and Picone, 1995) in Fig. 5 are low and near 1% for all conditions, even at the lowest SNR of 10 dB. Machine error rates are roughly ten times higher than those of humans, and increase at the lower SNRs. The low error rates reported in this study with humans are consistent with past studies (Kryter, 1960; Williams and Hecker, 1968) where humans recognized words in sentences with error rates below 2% in additive speech-shaped noise at SNRs ranging from 0 dB to 10 dB.

Similar low error rates for humans were found in a separate study which compared human and machine performance in quiet using *Wall Street Journal* sentences (Van Leeuwen et al., 1995). Eighty *Wall Street Journal* sentences were read by 23 speakers including both native USA and British speakers. These sentences were presented in quiet to 20 native USA and British listeners and to 10 non-native Dutch listeners. Although listeners were permitted to listen to sentence segments more than once, this option was used infrequently (roughly 10% of the time). Sentences were edited using a spelling checker to correct obvious typing and spelling errors. Word error rates were 2.6% for native listeners and 7.4% for non-native listeners. For comparison, the average word error rate for three high-performance 20,000-word recognizers that took part in a recent European *Wall Street Journal* evaluation was 12.6% for the same sentences (Van Leeuwen et al., 1995). One additional system developed for use with native USA speakers was not tested with British speech, but provided a lower error rate of 5.2% on the 40 sentences spoken by native USA speakers. The average error rates of machines tested on the full USA and British sentences were thus roughly five times greater than human error rates, and the error rate of one high-performance USA recognizer tested on a subset of 40 USA sentences was twice that of humans.

The results of these two studies demonstrate that error rates of humans are much lower than those of machines in quiet, and that error rates of current recognizers increase substantially at noise levels which have little effect on human listeners. Results also demonstrate that noise adaptation algorithms can dramatically improve recognition performance when noise spectra and SNRs are stationary and known during training. Even with noise adaptation, however, the 12.8% machine error rate shown in Fig. 5 at an SNR of 10 dB is more than an order of magnitude greater than the 1% human error rate. Two previous studies with isolated digits instead of sentences also found that human error rates are low in quiet and do not degrade at low SNRs. One study with four listeners found that the digit error rate was almost perfect (less than 0.5% errors) both in quiet and at an SNR of -3 dB (Pols, 1982). A second study with two listeners found that the human digit error rate was less than 1% both in quiet and at an SNR of 0 dB (Varga and Steeneken, 1993). This second study evaluated the performance of talker independent digit recognizers both with and without noise adaptation. All machine recognizers provided error rates of roughly 2% in quiet. At an SNR of 0 dB, error rates increased to almost 100% without noise adaptation and to 40% with the best noise adaptation algorithm.

The complete NAB corpus contains materials from many business publications and includes open-vocabulary test conditions where the recognition vocabulary is unlimited (Kubala, 1995; Stern, 1996). Large vocabulary HMM speech recognizers were recently compared to humans in one condition (Hub-3) of the 1995 ARPA continuous speech recognition evaluation. Speech was recorded using four different microphones in a slightly reverberant office environment with background acoustic noise which resulted in SNRs of roughly 20 dB. Speech materials were recorded using the same close-talking microphone used for the Resource Management and *Wall Street Journal* evaluations, with two high-quality cardioid studio microphones, and with a low-cost omni-directional electret microphone. Human error rates were measured with 15 listeners as described in (Deshmukh et al., 1996) using procedures similar to those used in (Ebel and Picone, 1995). Spellings of surnames and proper nouns were corrected and error rates were obtained from individual listeners and also using a majority vote from a committee of three listeners. Averaged over all microphone conditions, human error rates were 1.9% for individual listeners and 0.5% for the committee. The difference in this study between committee and individual error rates is twice as large as the difference found during the Spoke-10 evaluations (Ebel and Picone, 1995). This suggests that these Hub-3 materials are more difficult and that the committee results are more representative of the lowest human error rates that can be obtained by highly motivated listeners. Committee results are thus used for this corpus when comparing human and machine performance.

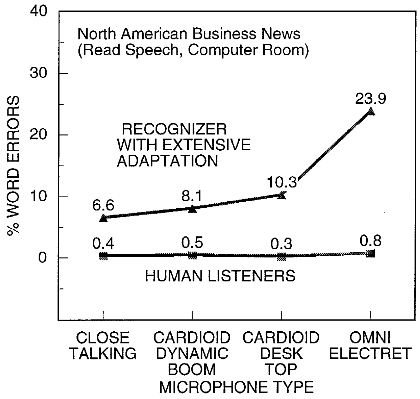

Fig. 6. Performance of human committee (Deshmukh et al., 1996) and of a high-performance HMM recognizer with channel adaptation (Woodland et al., 1996) with multiple microphones for the NAB corpus.

Fig. 6 shows human committee error rates from (Deshmukh et al., 1996) and machine error rates for the 65,000 word HMM recognizer that obtained the best results for this task (Woodland, et al. 1996). Human error rates are below 1.0% for all conditions. These error rates increase from roughly 0.4% with the three higher-quality microphones to 0.8% with the low-cost omni-directional electret microphone. Machine error rates increase from 6.6% with the close talking microphone to roughly 24% with the electret microphone. This increase occurs despite extensive adaptation algorithms which were used to compensate for microphone variability. Machine error rates under all conditions are again more than an order of magnitude greater than human error rates. Error rates degrade for both humans and machines with the omni-electret microphone. Under this more difficult condition, human error rates remain below 1.0% while machine error rates rise to roughly 24%.

4.4. Switchboard continuous speech recognition

Credit Card and other Switchboard telephone conversations have been used to evaluate the performance of large-vocabulary continuous-speech recognizers with spontaneous speech. Human performance on these materials can be estimated by evaluating the accuracy of transcriptions created by individual transcribers. Transcriptions were created by court reporters and temporary employees, and include non-speech sounds such as coughs, laughs and breath noise (LDC, 1995). The accuracy of over 14,000 transcribed words was carefully validated by linguists and speech scientists using repeated listening to both sides of each conversation, examinations of the waveform, and group-vote consensus decisions for difficult utterances which were faint, spoken rapidly, or partly masked by simultaneous speech from the opposite talker. The average transcription error rate, counting insertions, deletions and substitutions, was 4% (Martin, 1996). Speech recognition error rates on this corpus were initially high and near 67% as reported in (Young et al., 1994) for an HMM recognizer that provided state-of-the-art performance on the *Wall Street Journal* segment of the NAB corpus. More recent work has reduced these high rates to roughly 40% by using more training data and employing advanced adaptation and speech modeling techniques (e.g. Liu et al., 1996; Peskin et al., 1996).

Fig. 7 presents word error rates for an HMM recognizer described in (Liu et al., 1996) and for human transcribers on the spontaneous speech Switchboard corpus. The 43% error rate of the HMM recognizer is extremely high compared to the much lower machine rates obtained with the Resource Management and *Wall Street Journal* corpora using read speech. This increase in error rate is caused by many factors including the talking style, weak grammar used for conversations, and the limited bandwidth provided by telephone channels. Recognition experiments performed by Mitch Weintraub and reported in (Culhane, 1996) explored the importance of some of these factors. In one experiment, talkers engaged in spontaneous Switchboard-like conversations and then read transcriptions of these conversations. The machine recognition error rate was 52.6%

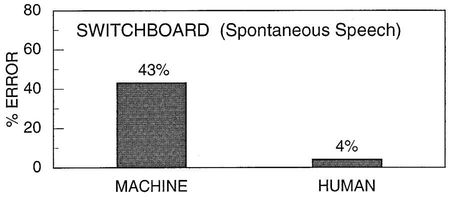

Fig. 7. Word error rates for humans and a high-performance HMM recognizer on phrases extracted from spontaneous telephone conversations in the Switchboard speech corpus (Liu et al., 1996; Martin, 1996).

for spontaneous speech and 28.8% for read versions of the same materials. This result demonstrates that the speaking style used with spontaneous speech is one of the primary reasons that error rates are high on the Switchboard corpus. Results with the Switchboard corpus thus demonstrate that error rates for machines can increase dramatically for spontaneous speech conditions where human performance remains high. Once again, the 4% error rate of humans is more than an order of magnitude less than the 43% error rate of a high-performance HMM recognizer.

4.5. Switchboard wordspotting

Machine wordspotters have been used to detect occurrences of 20 keywords in the Credit Card component of the Switchboard speech corpus. Evaluation of human wordspotting performance on the Credit Card corpus would involve listening to hours of spontaneous conversations searching for 20 keywords. This is too demanding for most listeners, and was replaced by a simpler discrimination task as described in (Chang and Lippmann, 1996). Two listeners focused on one keyword at a time and were randomly presented speech segments containing either a true keyword occurrence or a false alarm. After listening to a segment, subjects made a binary decision which indicated whether a specified keyword was contained in the segment. False alarms were generated by a high-performance hybrid HMM/neural-network wordspotter described in (Chang and Lippmann, 1996) which provides the best performance reported for a whole-word wordspotter on this corpus. Two listening conditions were used to evaluate the importance of context. In a ‘‘no context’’ condition, listeners were presented false alarms and keyword occurrences with 100 msec of extra speech both before and after each segment. This small amount of extra context was necessary to keep from chopping off the beginnings and end of keywords and almost never allowed users to hear adjacent words. In the ‘‘2 second context’’ condition, all sounds beginning 2 seconds before and ending 2 seconds after each utterance were presented. This context typically included most of the phrase containing each utterance. The resulting human judgements were used to compute new average detection and miss rates for each word by eliminating all utterances judged to be non-keywords. The highest human detection rate that can be achieved in this experiment is not 100% because humans did not listen to all wordspotter false alarms and thus could not eliminate all false alarms. The highest detection accuracy is 92.8% and the lowest miss rate is thus 7.2%.

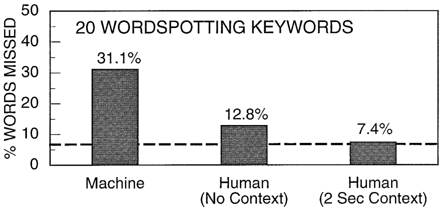

Fig. 8. Average miss rate for 20 keywords in the Credit Card component of the Switchboard corpus for a high-performance whole-word wordspotter and human listeners (Chang and Lippmann, 1996). The dotted line is the lowest miss rate possible when humans accept all correct keywords and reject all high-scoring false alarms.

Fig. 8 shows that human judgements reduce the average miss rate for keywords from 31.1% for the wordspotter to 12.8% for humans with no contextual information and 7.4% for humans with 2 seconds of contextual information. Without context, human judgements reduced the miss rate by roughly 18 percentage points. This substantial improvement in performance demonstrates that humans can make fine phonetic distinctions between similar sounding words without relying on surrounding semantic, syntactic or acoustic context. In these experiments, most false alarms were caused by words with strong vowel sounds that are similar to vowels in the keywords. For example the words ‘‘*car*’’ and ‘‘*far*’’ often caused false alarms for the keyword ‘‘*card*’’, and the word ‘‘*me*’’ caused false alarms for the keyword ‘‘*visa*’’. The ability of listeners to perceive differences between consonants in these similar sounding words was high, but not perfect, without surrounding contextual information.

Listeners made almost no errors when asked to determine whether a speech segment contained a specified keyword when they were provided 2 seconds of context before and after each extracted keyword. The average human judgement error rate was only 0.3% (8 out of 2314 judgements) with 2 seconds of context. This is indicated in Fig. 8 on the right by the small difference between the 7.4% miss rate obtained using human judgements and the lowest possible miss rate of 7.2% shown by the dotted line.

The results for this discrimination task are similar to results obtained from an older study which explored the intelligibility of speech fragments extracted from continuous speech (Pollack and Pickett, 1963). This study found that speech intelligibility for words in segments extracted from spontaneous speech is high (error rate roughly 5%) when the total duration of a segment is 2 seconds. Conversations on the subject of ‘‘college life’’ were surreptitiously recorded from four female student employees in a quiet environment, and fluent 2-second segments with no non-phonemic silences were extracted from these conversations. Ten to twelve segments were obtained for each talker containing from 4 to 15 successive words each. An example of one segment is ‘‘second or third but then I'd like to teach’’. Segments were played to crews of 22 to 29 listeners who were first presented only the first word in segment, then the first two words, then additional successive words until all words in a segment were presented. The average error rate decreased from roughly 45% for segments of 0.3 second duration, to 15% at 1 second duration, to 5% at 2 seconds duration. Results from this study and from (Chang and Lippmann, 1996) suggest that 2 seconds of speech provides sufficient context to compensate for differences in talkers, talking rate, channel characteristics, and linguistic context and also to provide coarticulatory acoustic information that may be spread across adjacent words.

5. Summary and discussion

Dramatic advances have recently been made in speech recognition technology. Large-vocabulary talker-independent recognizers provide error rates that are less than 10% for read sentences recorded in a quiet environment. Machine performance, however, deteriorates dramatically under degraded conditions. For example, error rates increase to roughly 40% for spontaneous speech and to 23% with channel variability and noise. Human error rates remain below 5% in quiet and under similar degraded conditions. Comparisons using many speech corpora demonstrate that human word error rates are often more than an order of magnitude lower than those of current recognizers in both quiet and degraded environments. In general, the superiority of human performance increases in noise, and for more difficult speech material such as spontaneous speech. Although current speech recognition technology is well suited to many practical commercial applications, these results suggest that there is much room for improvement.

Comparisons between human and machine error rates suggest the need for more fundamental research to improve machine recognition performance. This research could focus on four areas where past studies demonstrate the most dramatic differences between human and machine performance. First, results obtained with limited context suggest that human listeners perform more accurate low-level acoustic-phonetic modeling than machines. We can accurately recognize isolated digit sequences and spoken letters, we can recognize short segments extracted from spontaneous conversations, and we can accurately recognize words in nonsense sentences that provide little contextual information. These results suggest that one important direction for future research with machine recognizers is to improve low-level acoustic phonetic analysis.

Second, human recognition results obtained with channel variability and noise demonstrate that we can easily recognize speech with normally occurring degradations. Past studies have also demonstrated that we can understand speech with no training when highly unnatural distortions are applied, such as extreme waveform clipping (Licklider and Pollack, 1948), severe band-reject filtering (Lippmann, 1996), and extremely erratic linear frequency responses (Kryter, 1960). Machine recognition performance, however, often degrades dramatically with channel variability and noise. Machine recognizers typically provide best performance with degraded speech only when they are trained using degraded speech materials (e.g. Lippmann and Martin, 1987) or when internal parameters are adapted to mimic this type of training (e.g. Gopinath et al., 1995). These results suggest that a second important direction for future machine recognition research is to develop improved channel and noise adaptation algorithms which adapt rapidly to time varying noise and channel characteristics. These algorithms should maintain high recognition accuracy but require little a priori information concerning noise and channel characteristics.

Third, the extremely high error rates obtained with machines for spontaneous speech suggest that much further work should explore the differences between spontaneous and read speech and develop recognition approaches that provide good performance for this more difficult material.

Finally, the excellent human performance obtained when short 2–4 second segments of speech are extracted from spontaneous speech (e.g. Chang and Lippmann, 1996; Pollack and Pickett, 1963) suggest that further work is required on language modeling and on algorithms which can rapidly adapt to talker, channel and talker-style variability using only short speech segments. Two additional experiments with humans also suggest that rapid talker adaptation and improved grammar models are feasible. The first is an innovative experiment performed by Kakehi in Japan who demonstrated that human performance (the syllable error rate) degrades only slightly after changing talker, and that this small degradation is eliminated after hearing only three syllables (Kakehi, 1992). Four subjects listened in noise to 100-item lists formed from 100 common Japanese monosyllables containing 26 consonants and 5 vowels and produced by each of 100 different talkers. In one condition, all syllables in a list were produced by one talker, and listeners thus could adapt to this talker. In a second condition, the talker was varied at random for each syllable in a list and listeners were thus prevented from adapting to each talker. Syllable error rates increased slightly by 6 percentage points from roughly 24% to 30% when users could not adapt separately to each talker. Additional experiments measured the error rates of syllables immediately after a talker change with from 1 to 5 sequential presentations of syllables from a new talker. Results from these experiments demonstrate that only three syllables are required to adapt to a new talker and produce performance similar to that provided by a single-talker list. A second informal experiment explored the language modeling capabilities of humans using an interactive computer game called the ‘‘Shannon Game’’ where humans guess the next word in grammatical sentences (Jelinek, 1985). Human guesses were compared to trigram language models used in high-performance recognizers. It was found that humans beat the trigram model by factors of 3 or more in perplexity. This suggests that it should be possible to improve recognition accuracy substantially on continuous-speech tasks by producing language models with perplexities that are 1/3 those of current trigram grammars.

In addition to weaknesses in the above four areas, current machine recognizers still lack two fundamental capabilities that are essential for effective human-like performance. Modern recognizers cannot identify and learn new words and they cannot distinguish non-speech environmental sounds from acceptable speech input. Even human children can distinguish environmental sounds from speech and are able to distinguish new from known words. It is estimated that a high school graduate knows more than 160,000 words, counting roots, derivatives and compounds (e.g. Miller, 1991). Achieving this vocabulary requires learning roughly 10 new words a day. Few machine recognizers have the capability of distinguishing known from unknown words, and even fewer can automatically learn the meanings of new words. Machine recognizers also cannot currently be used in normal environments without a close-talking or directional microphone or without using a push-to-talk switch because desired speech inputs cannot be separated from other environmental sounds. Common transient and intermittent environmental sounds may be interpreted by many modern high-performance recognizers as well-formed sentences. Some preliminary work has been performed in these two areas (e.g. Gorin et al., 1994; Brown and Cooke, 1994) but few objective evaluations have been performed, and algorithms are still in the early developmental stage.

Many researchers are already exploring approaches to improve acoustic-phonetic modeling, to lower error rates in noise and with channel variability, to improve performance with spontaneous speech, to improve language modeling, and to provide more rapid talker, noise and channel adaptation (e.g. Young, 1996; Bourlard et al., 1996). The studies on human perception described in this paper provide baseline error rates to evaluate these new algorithms and an existence proof that accurate speech recognition is possible with degraded speech and spontaneous speaking styles. Further studies comparing human and machine performance will be required to update baseline human results on new tasks and to monitor improvements in speech recognition technology. Further human studies should also be designed to provide hints concerning the algorithms used by biological computation to provide high performance when inputs include both speech and nonspeech sounds and with the types of talker and acoustic variability encountered in everyday environments.

Acknowledgements

The author would like to thank the three reviewers for their thoughtful evaluations and comments. This work was sponsored by DARPA. Opinions, interpretations, conclusions, and recommendations are those of the author and are not necessarily endorsed by the United States Air Force.

References

Bourlard, H., Hermansky, H., Morgan, N., 1996. Towards increasing speech recognition error rates. Speech Communication 18 (3), 205–231.

Brown, G.J., Cooke, M.P., 1994. Computational auditory scene analysis. Computer Speech and Language 8, 297–336.

Chang, E., Lippmann, R., 1996. Improving wordspotting performance with artificially generated data. Proc. IEEE Internat. Conf. Acoust. Speech Signal Process., pp. 526–529.

Chou, W., Lee, C.H., Juang, B.H., 1994. Minimum error rate training of inter-word context dependent acoustic model units in speech recognition. Proc. Internat. Conf. on Spoken Language Processing, Yokohama, Japan, pp. S09:3.1–3.4.

Cole, R., Fanty, M., Muthusamy, Y., Gopalakrishnan, M., 1990. Speaker-independent recognition of spoken english letters. Proc. 1990 IEEE INNS Internat. Joint Conf. on Neural Networks, San Diego, CA, Vol. 2, pp. 45–51.

Culhane, C., 1996. Summary of session 7 – Conversational and multi-lingual speech recognition. Proc. DARPA Speech Recognition Workshop. Morgan Kaufmann, Harriman, NY, pp. 143–144.

Daly, N., 1987. Recognition of words from their spellings: Integration of multiple knowledge sources. Master's Thesis, Massachusetts Institute of Technology.

Deshmukh, N., Ganapathiraju, A., Duncan, R.J., Picone, J., 1996. Human speech recognition performance on the 1995 CSR Hub-3 corpus. Proc. DARPA Speech Recognition Workshop. Morgan Kaufmann, Harriman, NY, pp. 129–134.

Ebel, W.J., Picone, J., 1995. Human speech recognition performance on the 1994 CSR Spoke 10 corpus. Proc. Spoken Language Systems Technology Workshop. Morgan Kaufmann, Austin, TX, pp. 53–59.

Fletcher, H., Steinberg, J.C., 1929. Articulation testing methods. Bell System Technical J. 8, 806–854.

Godfrey, J.J., Holliman, E.C., McDaniel, J., 1992. SWITCHBOARD: Telephone speech corpus for research and development. Proc. IEEE Internat. Conf. Acoust. Speech Signal Process., pp. 517–520.

Gopinath, R.A., Gales, M., Gopalakrishnan, P.S., Balakrishnan-Aiyer, S., Picheny, M.A., 1995. Robust speech recognition in noise – Performance of the IBM Continuous Speech Recognizer on the ARPA noise spoke task. Proc. Spoken Language Systems Technology Workshop. Morgan Kaufmann, Austin, TX, pp. 127–130.

Gorin, A.L., Levinson, S.E., Sankar, A., 1994. An experiment in spoken language acquisition. IEEE Trans. Speech Audio Process. 2, 224–240.

Huang, X.D., Lee, K.F., Hon, H.W., Hwang, M.Y., 1991. Improved acoustic modeling with the SPHINX Speech Recognition System. Proc. IEEE Internat. Conf. Acoust. Speech Signal Process., pp. 345–348.

Jelinek, F., 1985. The development of an experimental discrete dictation recognizer. Proc. IEEE 73, 1616–1624.

Kakehi, K., 1992. Adaptability to differences between talkers in Japanese monosyllabic perception. In: Tohkura, Y., Vatikiotis-Bateson, E., Sagisaka, Y. (Eds.), Speech Perception, Production and Linguistic Structure. IOS Press, Amsterdam, pp. 135–142.

Kryter, K.D., 1960. Speech bandwidth compression through spectrum selection. J. Acoust. Soc. Amer. 32, 547–556.

Kubala, F., 1995. Design of the 1994 CSR Benchmark Tests. Proc. Spoken Language Systems Technology Workshop. Morgan Kaufmann, Austin, TX, pp. 41–46.

LDC, 1995. SWITCHBOARD: A User's Manual, Catalog Number LDC94S7. Linguistic Data Consortium, University of Pennsylvania, Philadelphia, PA.

Lee, K.F., 1989. Automatic Speech Recognition. Kluwer, Boston, MA.

Leonard, R.G., 1984. A database for speaker-independent digit recognition, Proc. IEEE Internat. Conf. Acoust. Speech Signal Process., pp. 42.11.1–42.11.4.

Licklider, J.C.R., Pollack, I., 1948. Effects of differentiation, integration, and infinite peak clipping upon the intelligibility of speech. J. Acoust. Soc. Amer. 20, 42–51.

Lippmann, R.P., 1996. Accurate consonant perception without mid-frequency speech energy. IEEE Trans. Speech Audio Process. 4, 66–69.

Lippmann, R.P., Martin, E.A., 1987. Multi-style training for robust isolated-word speech recognition. Proc. IEEE Internat. Conf. Acoust. Speech Signal Process., pp. 705–708.

Lippmann, R.P., Braida, L.D., Durlach, N.I., 1981. A study of multichannel amplitude compression and linear amplification for persons with sensorineural hearing loss. J. Acoust. Soc. Amer. 69, 524–534.

Liu, F.-H., Picheny, M., Srinivasa, P., Monkowski, M., Chen, J.,

1996. Speech recognition on Mandarin Call Home: A large-vocabulary, conversational, and telephone speech corpus. Proc. IEEE Internat. Conf. Acoust. Speech Signal Process., pp. 157–160.

A. Martin, 1996. Personal communication.

Miller, G.A., 1962. Decision units in the perception of speech. Institute of Radio Engineers Transactions on Information Theory 8, 81–83.

Miller, G.A., 1991. The Science of Words. Freeman, New York.

Pallett, D.S., 1991. DARPA resource management and ATIS benchmark test poster session. Proc. DARPA Speech and Natural Language Workshop. Morgan Kaufmann, Austin, TX, pp. 49–58.

Pallett, D.S., Fiscus, J.G., et al., 1995. 1994 benchmark tests for the ARPA Spoken Language Program. Proc. Spoken Language Systems Technology Workshop. Morgan Kaufmann, Austin, TX, pp. 5–36.

Paul, D., Baker, J., 1992. The design for the Wall Street Journal-based CSR corpus. Proc. DARPA Speech and Natural Language Workshop. Morgan Kaufmann, Austin, TX, pp. 357–360.

Peskin, B., Connolly, S., Gillick, L., Lowe, S., McAllaster, D., Nagesha, V., Van Mulbregt, P., Wegmann, S., 1996. Improvements in Switchboard recognition and topic identification. Proc. IEEE Internat. Conf. Acoust. Speech Signal Process., pp. 303–306.

Pollack, I., Pickett, J.M., 1963. The intelligibility of excerpts from conversation. Language and Speech 6, 165–171.

Pols, L.C.W., 1982. How humans perform on a connected-digits data base. Proc. IEEE Internat. Conf. Acoust. Speech Signal Process., pp. 867–870.

Price, P., Fisher, W.M., Bernstein, J., Pallett, D.S., 1988. The DARPA 1000-word resource management database for continuous speech recognition. Proc. IEEE Internat. Conf. Acoust. Speech Signal Process., pp. 651–654.

Stern, R.M., 1996. Specification of the 1995 ARPA Hub 3 evaluation: Unlimited vocabulary NAB news baseline. Proc. Speech Recognition Workshop. Morgan Kaufmann, Harriman, NY, pp. 5–7.

Van Leeuwen, D.A., Van den Berg, L.G., Steeneken, H.J.M., 1995. Human benchmarks for speaker independent large vocabulary recognition performance. Eurospeech, Madrid, pp. 1461–1464.

Varga, A., Steeneken, H.J.M., 1993. Assessment for automatic speech recognition: II. NOISEX-92: A database and an experiment to study the effect of additive noise on speech recognition systems. Speech Communication 12 (3), 247–251.

Williams, C.E., Hecker, M.H.L., 1968. Relation between intelligibility scores for four test methods and three types of speech distortion. J. Acoust. Soc. Amer. 44 (4), 1002–1006.

Woodland, P.C., Gales, M.J.F., Pye, D., Valtchev, V., 1996. The HTK large vocabulary recognition system for the 1995 ARPA H3 Task. Proc. Speech Recognition Workshop. Morgan Kaufmann, Harriman, NY, pp. 99–104.

Young, S.J., 1996. A review of large-vocabulary continuous-speech recognition. IEEE Signal Process. Mag. 13, 45–57.

Young, S.J., Woodland, P.C., Byrne, W.J., 1994. Spontaneous speech recognition for the credit card corpus using the HTK toolkit. IEEE Trans. Speech Audio Process. 2 (4), 615–621.

Zue, V., Glass, J., Phillips, M., Seneff, S., 1989. The MIT SUMMIT speech recognition system: A progress report. Proc. DARPA Speech and Natural Language Workshop. Morgan Kaufmann, Philadelphia, PA, pp. 179–189.