LINGUISTIC DISSECTION OF SWITCHBOARD-CORPUS AUTOMATIC SPEECH RECOGNITION SYSTEMS

Steven Greenberg and Shuangyu Chang

International Computer Science Institute
1947 Center Street, Berkeley, CA 94704, USA
{steveng,shawnc}@icsi.berkeley.edu

ABSTRACT

A diagnostic evaluation of eight Switchboard-corpus recognition systems was conducted in order to ascertain whether word-error patterns are attributable to a specific set of linguistic factors. Each recognition system's output was converted to a common format and scored relative to a reference transcript derived from phonetically hand-labeled data. This reference material was analyzed with respect to ca. forty acoustic, linguistic and speaker characteristics, which in turn, were correlated with recognition-error patterns via decision-trees and other forms of statistical analysis. The most consistent factors associated with superior recognition performance pertain to accurate classification of phonetic segments and articulatory-acoustic features. Other factors correlated with word recognition are syllable structure, prosodic stress and speaking rate (in terms of syllables per second).

1. INTRODUCTION

The present study represents an effort to disect the functional architecture of large-vocabulary speech recognition systems used in the annual NIST-sponsored Switchboard Corpus evaluation. The Switchboard corpus [5] has been used in recent years (in tandem with the Call Home and Broadcast News corpora) to assess the state of automatic speech recognition (ASR) for spoken English. Switchboard is unique among the large-vocabulary corpora in having a substantial amount of material that has been phonetically labeled and segmented by linguistically trained individuals (Switchboard Transcription Project - http://www.icsi.berkeley.edu/real/stp [6] [7]) and thus provides a crucial set of "reference" materials with which to assess and evaluate the phonetic and lexical classification capabilities of current-generation ASR systems. In a separate paper [8] we describe in detail the methods used to evaluate the Switchboard recognition systems developed by eight separate sites (cf. Figures 1-3), as well as delineating a few key macroscopic analyses of the diagnostic material (cf. Figures 4-6). In that earlier study we concluded that the properties most predictive of lexical recognition in ASR systems pertain to the accuracy of *phonetic* classification (both at the phone and articulatory-feature level - cf. Figure 4 and Section 6). However, this conclusion is based primarily on the results of a decision-tree analysis (cf. Section 6) that may be of too coarse a nature to reveal certain linguistic patterns germane to ASR performance. For this reason, the current study examines the contribution of such factors as syllable structure, prosodic stress and speaking rate, as well as those pertaining to phonetic and articulatory-feature classification, in order to identify those linguistic parameters that are most highly correlated with word-recognition performance among the eight ASR systems.

2. CORPUS MATERIALS

The evaluation was performed on a fifty-four-minute, phonetically annotated subset of the Switchboard corpus (http://www.icsi.berkeley.edu/real/phoneval). The material had previously been manually segmented at the syllabic and lexical levels and was segmented into phonetic segments using an automatic procedure trained on a seventy-two-minute subset of the phonetic-transcription material that had been previously segmented by hand. The resulting segmentation was verified using a combination of manual and automatic methods.

3. EVALUATION FORMAT

Eight separate sites participated in the evaluation - AT&T, BBN, Cambridge University (CU), Dragon Systems (DRAG), Johns Hopkins University (JHU), Mississippi State University (MSU), SRI International and the University of Washington (UW). Each site was asked to submit two different sets of material:

(1) the word and phonetic-segment output of the recognition system used for competitive (i.e., non-diagnostic) portion of Switchboard, and

(2) the word and phone-level output of forced-alignments associated with the same material (provided by six sites).

In order to score the submissions in terms of phone-segments and words correct, as well as perform detailed analyses of the error patterns, it was necessary to convert the submissions into a common format. This required that:

(1) each site's phonetic symbol set be mapped onto a common reference similar to that used to phonetically annotate the Switchboard corpus (STP). Care was taken to insure that the mapping was conservative in order that a site not be penalized for using a symbol set distinct from STP. In addition, phonetic symbols not contained in a site's inventory were mapped to the more fine-grained STP phone set (http:www.icsi.berkeley.edu/real/phoneval).

(2) a reference set of materials at the word, syllable and phone levels was created in order to score the material submitted. This reference material included:
   (a) word-to-phone mapping
   (b) syllable-to-phone mapping
   (c) word-to-syllable mapping
   (d) time points for the phones and words in the reference materials

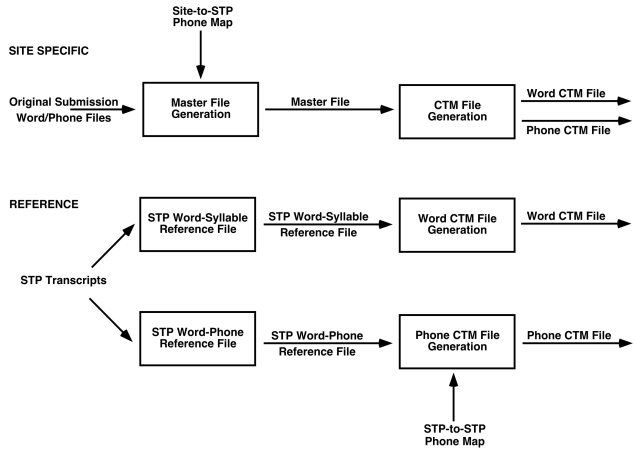

Figure 1: The initial phase of the diagnostic evaluation. Materials submitted by each site are converted into a format designed for scoring (CTM files) relative to the reference transcript (at the phonetic, syllable and word level). From [8]

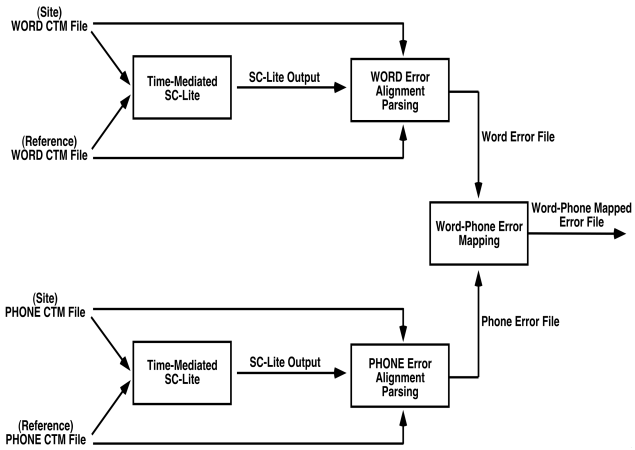

Figure 2: The evaluation's second phase involves time-mediated scoring of both the word- and phone-level output of the recognition and forced-alignment materials. The scored output is used to compile summary tables. From [8]

(3) time-mediated synchronization of the phone and word output of the submission material with that of the reference set.

The conversion process (Figure 1) was required in order that the submissions be scored at the word and phonetic-segment levels using SC-Lite, a program developed at the National Institute of Standards and Technology (NIST) to score competitive ASR evaluation submissions.

4. SCORING THE RECOGNITION SYSTEMS

SC-Lite scores each word (and phone) in terms of being correct or not, as well as designating the error as one of three types - a substitution (i.e., *a -> b*), an insertion (*a -> a+b*) or a deletion (*a -> ø*). A fourth category, *null*, occurs when the error can not be clearly associated with one of the other three categories (and usually implies that the error is due to some form of formatting discrepancy).

For both phone- and word-scoring it was necessary to develop a method enabling each segment (either word or phone) in the submission to be unambiguously associated with a corresponding symbol (or set of symbols) in the reference material. This was accomplished by using time-mediated boundaries as synchronizing delimiters. Because the word and phone segmentation of the submission materials often deviate from those of the STP-based reference materials an algorithm was developed to minimize the time-alignment discrepancy.

Files (in NIST's CTM format) were generated for each site's submission (separate files for recognition and forced alignment) and this material processed along with the CTM files associated with the word- and phone-level reference material (Figure 2). The resulting output was used as the basis for generating the data contained in the summary tables ("big lists" - cf. [8]) and which form the foundation of the current analyses.

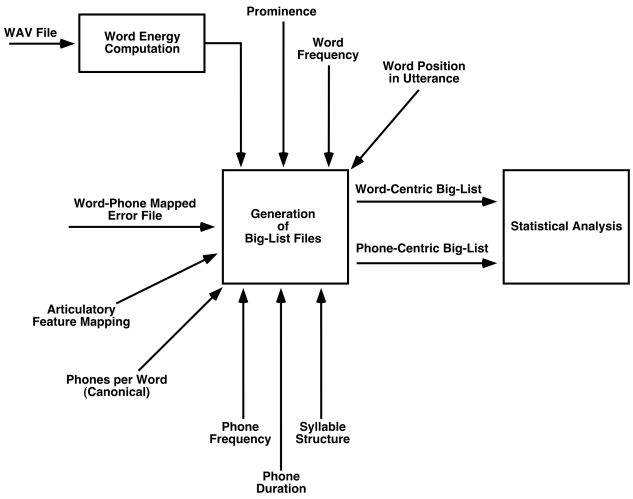

Figure 3: Phase three of the analysis consists of computing several dozen parameters associated with the phone- and word-level representations of the speech signal and compiling these into summary tables ("big lists"). A complete list of the parameters computed can be found in [8].

5. WORD AND PHONE ERROR PATTERNS

Word (Figure 4) and phone (Figures 4-6) level error patterns were computed for both the forced-alignment (based on word-level transcripts) and unconstrained recognition material. Phone recognition error is relatively high for forced recognition (35-49% - Figure 5) only slightly less than the phone classification error for unconstrained recognition (39-55% - Figure 6). The

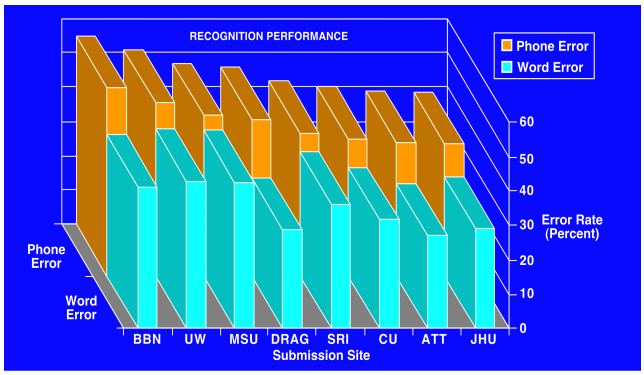

Figure 4: A comparison of the word and phonetic-segment error for the recognition component of the diagnostic Switchboard evaluation for all eight participating sites. From [8]

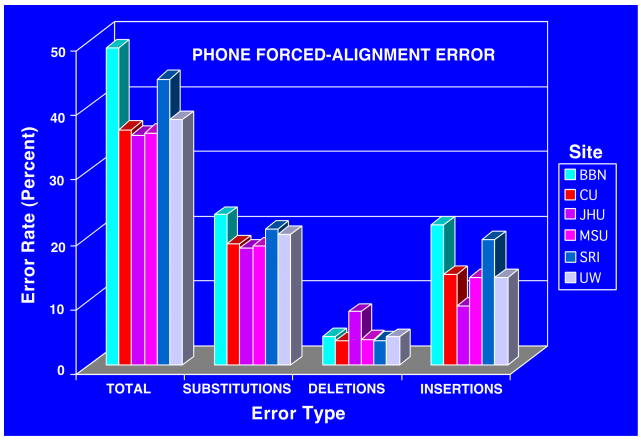

Figure 5: Phonetic-segment errors in the forced-alignment component of the Switchboard diagnostic evaluation for the six participating sites. From [8]

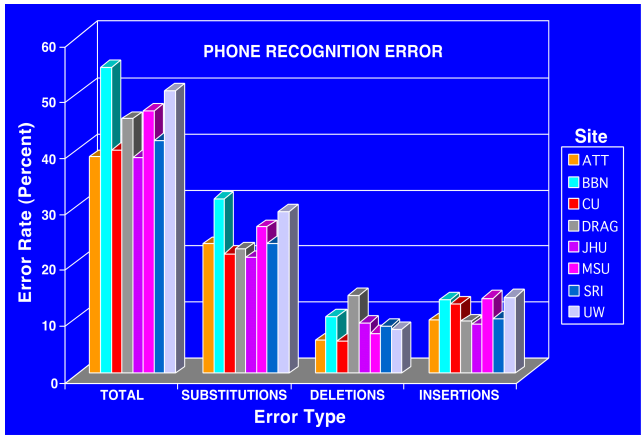

Figure 6: The percentage of phone errors for the recognition component of the Switchboard diagnostic evaluation. Data are from all eight participating sites. From [8]

difference in performance between the two conditions is much smaller than anticipated, suggesting that the ASR systems may not be optimized for classification of phonetic segments.

The word error rate for the ASR systems ranges between 27 and 43%, about 50% higher than that observed for the competitive portion of the evaluation [10]. The higher error rate is probably due to several factors. First, the diagnostic component of the evaluation contains relatively short utterances (mean duration = 4.76 sec) from hundreds of different speakers. In contrast, the competitive evaluation is composed of complete dialogues lasting ca. five minutes and produced by only forty different speakers [10]. Most (if not all) of the recognition systems normally use some form of speaker adaptation, which works most effectively over long spans of speech. Short utterances, such as those used in the diagnostic evaluation, are likely to mitigate the beneficial effect of speaker adaptation.

Figure 4 illustrates the relationship between phone- and word-error magnitude across submission sites. The correlation between the two *(r)* is 0.78, suggesting that word recognition may largely depend on the accuracy of recognition at the phonetic-segment level. Certain sites, such as AT&T and Dragon, deviate from this pattern in that their performance exhibits a lower word error than would be expected based solely on recognition at the phonetic-segment level. These systems may possess extremely good pronunciation models that partially compensate for the relative deficiencies of phone classification.

6. DECISION-TREE ANALYSIS OF ERRORS

In order to gain some initial insight into the factors governing word errors in recognition performance the STP-based, reference component of the Switchboard corpus was analyzed with respect to ca. forty separate parameters pertaining to speaker, linguistic and acoustic properties of the speech materials, including energy level, duration, stress pattern, syllable structure, speaking rate and so on (cf. [8] for a complete list of parameters). Decision trees [13] were used as a means of identifying the most important factors associated with word-error rate across sites. The error data were partitioned into four separate domains:

(1) Substitutions versus all other data (both correct and incorrect),
(2) Deletions versus all other data (both correct and incorrect),
(3) Substitution versus deletions (i.e., excluding words correctly recognized), and
(4) Substitutions versus insertions (i.e., excluding words correctly recognized)

Substitution Errors (versus All Else)

Half of the word errors involve substitutions (Figure 6), and it is therefore of interest to identify the parameters associated with this single most important component of recognition performance. For seven of the eight submissions the parameter dominating the decision tree at the highest (or second-highest) node-level is the number of phonetic-segment substitution errors within a word. The probability of a word being incorrectly recognized increases significantly when more than (an average of) ca. 1.5 phones are misclassified. Other important parameters in the decision trees are the acoustic-articulatory feature distance (AFDIST) between the correct and hypothesized word (which is also related to the probability of correct phone classification), (unigram) frequency of the reference word (WDFREQ), and whether the preceding (PREWDER) or following (PSTWDER) word is incorrectly recognized.

Deletions (versus All Else)

Deletions account for ca. 25% of the word errors. For all sites, the dominant factor associated with deletion errors pertains to either the number of phonetic segments correctly recognized (PHNCOR) or the acoustic-articulatory phonetic-feature distance between the reference and hypothesized word. Other important parameters are the number of phone insertions (PHNINS) and substitutions (PHSUB), word frequency and duration of the reference word (REFDUR).

Substitution versus Deletion Errors

Additional information concerning the source of word errors can be obtained by analyzing factors distinguishing different types of error. For distinguishing substitution from deletion errors two sets of parameters appear to be most important - phonetic-segment classification (PHNSUB,

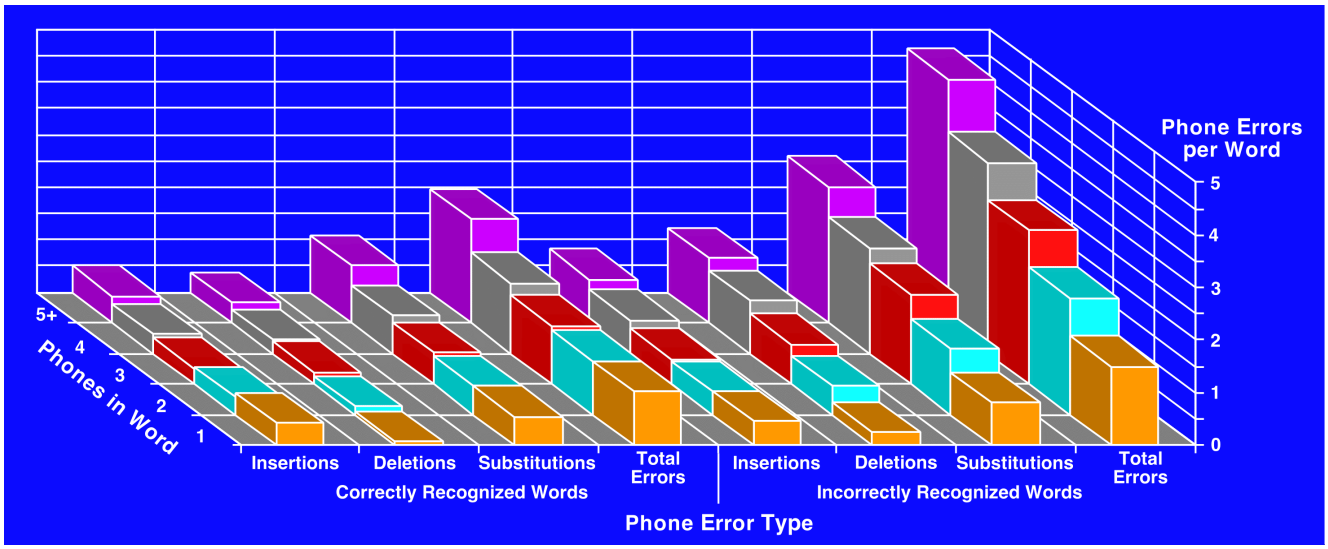

Figure 7: The number of phonetic-segment errors per word as a function of the number of phones contained in the word. The data are partitioned into two broad classes - correctly and incorrectly recognized words. Within each subset the data are divided into four classes, based on the type of error. Data represent averages across the eight ASR systems. Words of designated length "5+"contain five *or more* phones.

PHNINS, PHNCOR, AFDIST) and the duration of the reference (REFDUR) and hypothesized (HYPDUR) words.

Substitution versus Insertion Errors

The duration of the hypothesized word is the most important parameter distinguishing substitution from insertion errors, followed in importance by phonetic segment classification factors (PHNSUB, PHNINS, PHNDEL, AFDIST), frequency of occurrence of the phonetic segments in a word (PHNFREQ) and the error status of the preceding and following words (PREWDER, PSTWDER).

General Trends of the Decision Tree Error Analysis

The most important parameters associated with word-recognition error are those pertaining to the correct identification of a word's phonetic composition (at either the phone or articulatory-acoustic, phonetic-feature level). The results of the decision-tree analyses (cf. [8] for further detail) are also of interest because of the parameters that are *absent* from the decision trees - prosodic prominence, syllable structure and speaking rate. All of these parameters have been suggested as important factors associated with word-error rate. The decision trees suggest that such parameters do not account for word errors "across the board" in the way that acoustic-phonetic factors do.

7. PHONE ERROR PATTERNS IN WORDS

It is of interest to ascertain if the number of phonetic segments in a word bears any relation to the pattern of phone errors in both correctly and incorrectly recognized words (cf. Figure 7 and [3]). The "tolerance" for phonetic segment errors in correctly recognized words is not linearly related to the length of the word. The tolerance for error (ca. 1-1.5 phones) is roughly constant for word lengths of four phones or less. This pattern is observed regardless of the form of error. The relatively low tolerance for phone misclassification (except for words of very short length) implies that the pronunciation and language models possess only a limited capacity to compensate for errors at the phonetic-segment level.

In contrast, the average number of phones misclassified in incorrectly recognized words does increase in quasi-linear fashion as a function of word length (with the possible exception of insertions), a pattern consistant with the importance of phonetic classification for accurate word recognition.

8. ARTICULATORY-FEATURE ANALYSIS

Phonetic segments can be decomposed into more elementary constituents based on their articulatory bases, such as place (e.g., labial, labio-dental, alveolar, velar), manner (e.g., stop, fricative, affricate, nasal, liquid, glide, vocalic), voicing and lip-rounding. Two additional dimensions were also used in the current analyses - front-back articulation (vowels only) and the general distinction between consonantal and vocalic segmental forms.

There are approximately three times the number of

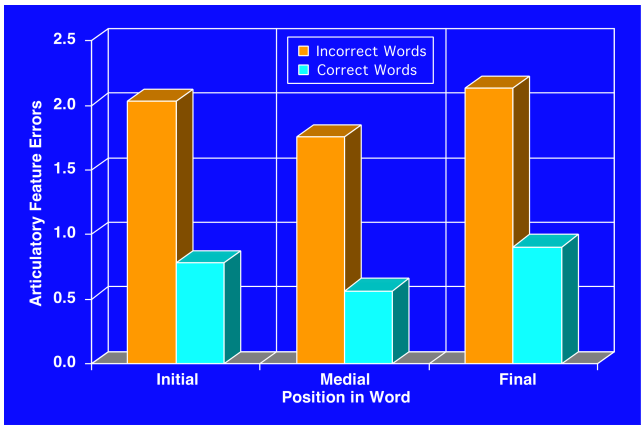

Figure 8: The average number of articulatory-feature errors per phone as a function of the phonetic segment's position within the word. The data are partitioned, depending on whether the word was correctly recognized or not. Data represent averages across the eight ASR systems.

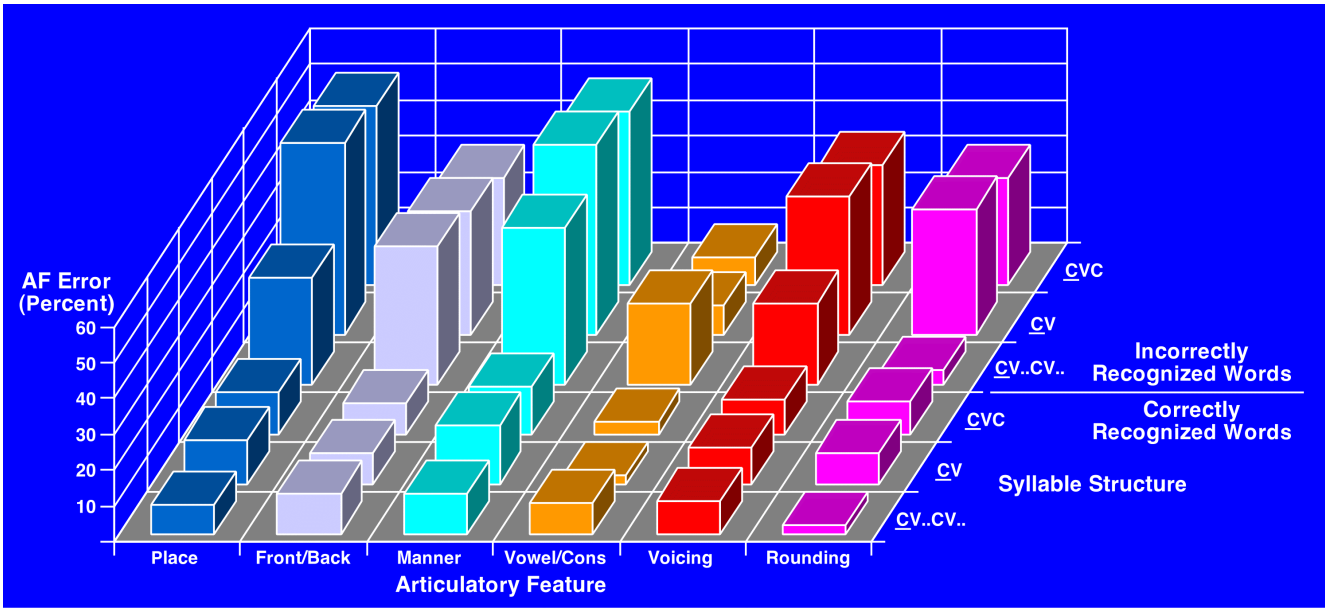

Figure 9: The average error in classification of articulatory features associated with each phonetic segment as a function of the position of the phone within the syllable - in this instance for ***consonantal onsets***. "CV..CV" indicates that the context was a polysyllabic word. Data represent averages across the eight ASR systems.

articulatory-feature (AF) errors (per phone) when the word is misrecognized, regardless of the position of the segment within the word (Figure 8). This contrast between correctly and incorrectly recognized words is maintained when the data are partitioned in articulatory-feature classes.

Because words are organized into a distinct set of syllabic forms in English it is of interest to partition the AF-error data according to lexical syllable structure (i.e., consonant and vowel sequences) and separately examine the pattern of errors for consonantal onsets, vocalic nuclei and consonantal codas. These data, illustrated in Figures 9-11, exhibit certain interesting patterns.

The articulatory features associated with consonantal onsets (Figure 9) exhibit a relatively low tolerance for error. Moreover, the error rate for is four to five times greater for AFs in misclassified words compared to correctly recognized lexical items. Place and manner features are particularly prone to error in misclassified words, suggesting that these onset-consonant AFs are particularly important for correctly distinguishing among words.

The AF error patterns for consonantal codas (Figure 10) are similar to those associated with consonantal onsets, except that there is a slightly higher (ca. 50%) tolerance for error among the former. Both onsets and codas also exhibit an interesting pattern with respect to polysyllabic words. Lip-rounding appears to play a much smaller role in these longer words while vowel/consonant confusions seem to be more pronounced in such contexts.

The AF error patterns associated with vocalic nuclei (Figure 11) display a pattern different from those associated with onsets and codas. There is a much higher tolerance of error in classification of AFs for the correctly

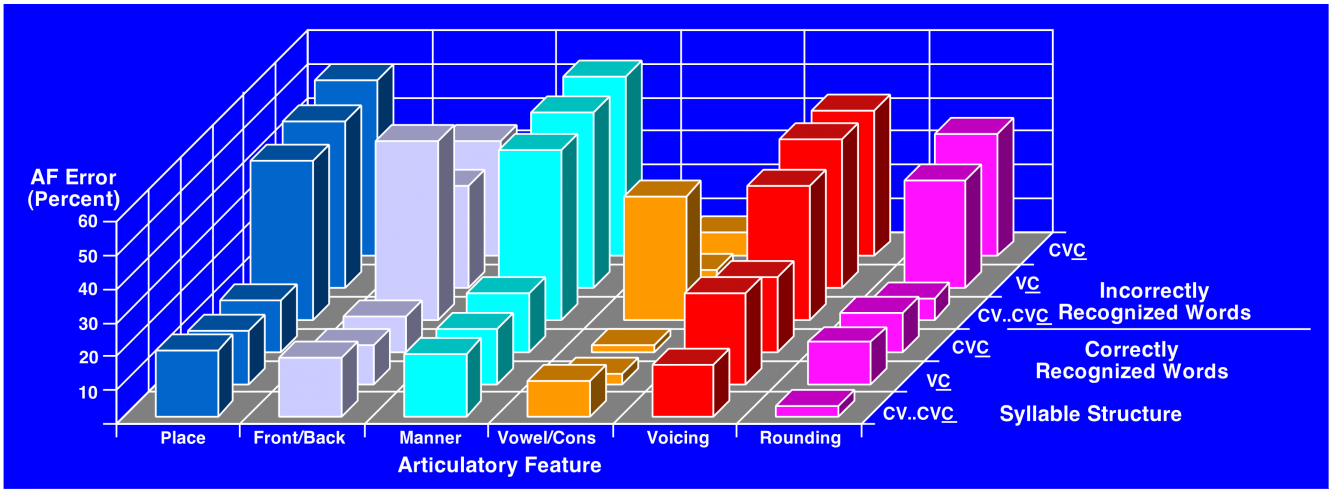

Figure 10: The average error in classification of articulatory features associated with each phonetic segment as a function of the position of the phone within the syllable - in this instance for ***consonantal codas***. "CV..CVC" indicates that the context was a polysyllabic word. Data represent averages across the eight ASR systems.

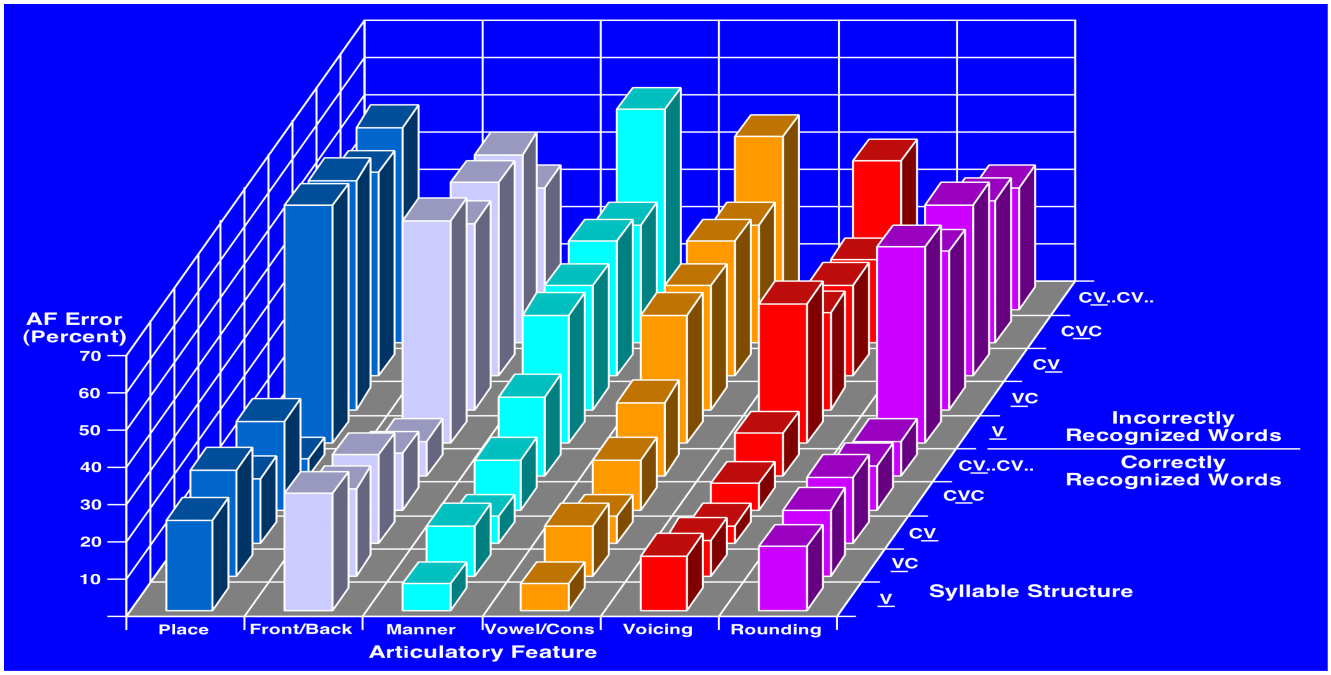

Figure 11: The average error in classification of articulatory features associated with each phonetic segment as a function of the position of the phone within the syllable - in this instance for ***vocalic nuclei***. "CV..CV" indicates that the context was a polysyllabic word. Data represent averages across the eight ASR systems.

recognized words - particularly marked for the place and front/back features. Moreover, there is a considerably higher degree of AF classification error among the nuclei compared to onsets and codas, particularly among the place and front-back dimensions. There is also an exceedingly high proportion of vowel/consonant confusions. Such data imply that classification of vocalic nuclei is considerably less precise than for the onsets and codas.

9. SYLLABLE STRUCTURE ANALYSIS

The AF classification patterns suggest that syllabic structure may help to provide a keener understanding of recognition errors. Figure 12 shows the word error as a function of syllable form. The highest error rates are associated with vowel-initial syllables. Syllables with complex (i.e., consonant cluster) onsets and codas tend to exhibit a relatively low word-error rate, as do polysyllabic words. This effect is particularly pronounced with respect to deletions, and vowel-initial syllables are particularly prone to such errors.

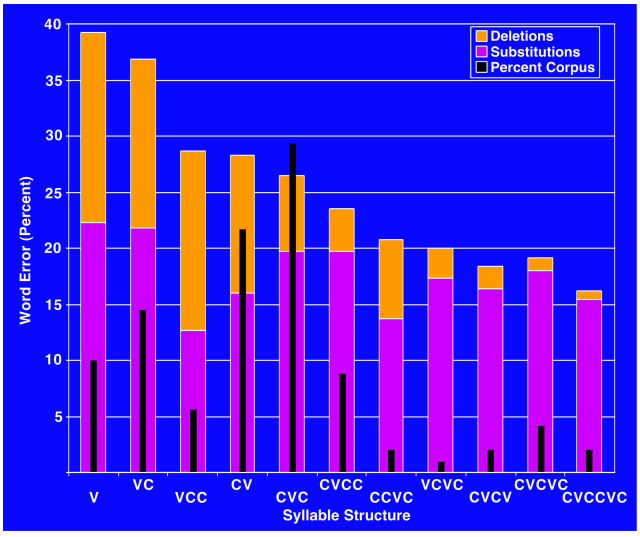

Figure 12: Word error as a function of lexical syllable form (C=consonant, V=vowel), partitioned into substitution and deletion errors. The proportion of the corpus associated with each syllabic form is also indicated (e.g, V forms are 10% of the total).

There is a certain degree of variability in the error patterns associated with syllable form across the eight ASR systems (Figure 13). Certain sites (such as AT&T, BBN and UW) exhibit a similar error rate across consonant-initial words. Other sites (e.g., Dragon, JHU and SRI) do particularly well on polysyllabic words. Virtually all sites show a pronounced increase in errors associated with vowel-initial words (AT&T is an exception).

The pattern of errors associated with syllable form suggest that future-generation ASR systems would benefit from an accurate means of parsing the acoustic signal into syllabic segments (cf. [14]) and reliably distinguishing vocalic constituents from their consonantal counterparts (cf. [2]).

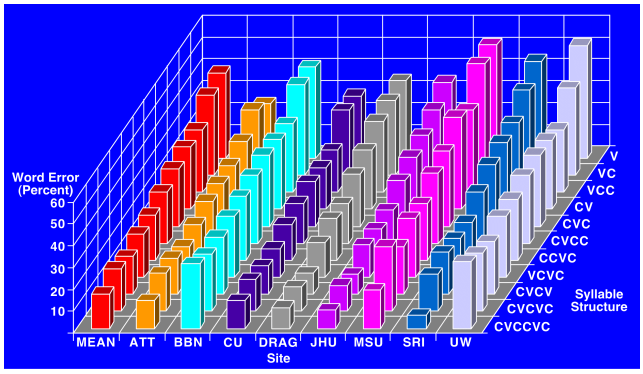

Figure 13: Word error as a function of syllable form for each of the eight ASR systems (and the mean, replotted from Figure 12).

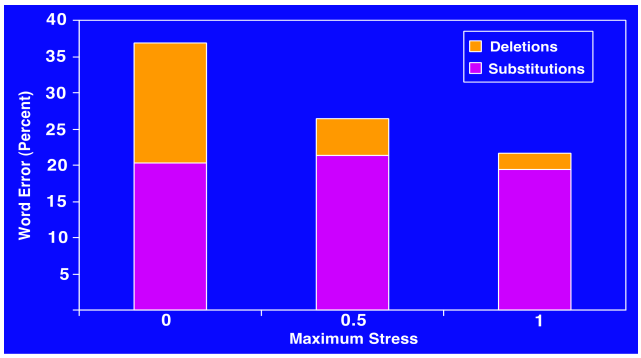

Figure 14: The average error rate for the eight ASR systems as a function of the maximum stress associated with a word. The data are partitioned into substitution and deletion errors. The total error is the sum of the two (insertions were excluded from the analysis). A maximum stress of "0" indicates that the word was completely unstressed. "1" indicates that at least one syllable in the word was fully stressed. An intermediate level of stress is associated with a value of 0.5.

10. THE EFFECT OF PROSODIC STRESS

Prosodic stress is an important means of providing informational emphasis in spontaneous speech [1] [9] and affects the pronunciation of phonetic elements [4] [6]. The diagnostic portion of the Switchboard corpus was prosodically labeled by two linguistically trained individuals and this material (http://www.icsi.berkeley.edu/~steveng/prosody) used to ascertain the relation between error rate and prosodic stress.

There is a 50% higher probability of a recognition error when a word is entirely unstressed (Figure 14). The relation between lexical stress and word-error rate is particularly apparent for deletions (Figure 14) and is manifest across all ASR systems (Figure 15). The relation between deletion errors and prosodic stress suggests that ASR systems may benefit from incorporating methods for automatic classification of acoustic prominence (cf. [15] [16]).

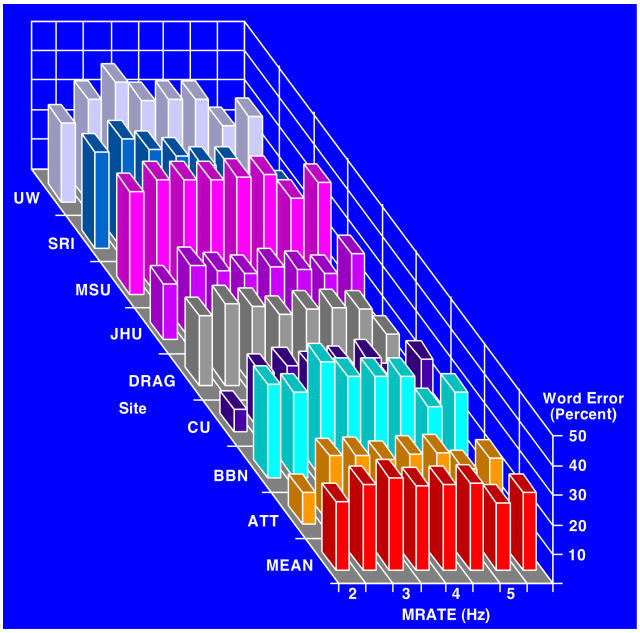

Figure 16: The relationship between word-error rate for each of the eight ASR systems (as well as the mean) and an *acoustic* measure of speaking rate (MRATE).

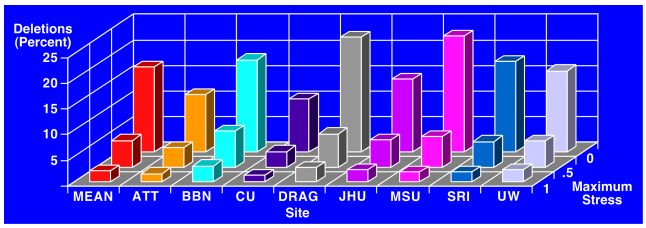

Figure 15: The average number of word deletions as a function of the maximum stress level associated with a word for each of the eight ASR systems.

11. THE EFFECT OF SPEAKING RATE

ASR systems generally have more difficulty recognizing speech that is of particularly fast [11] [12] or slow [12] tempo. A variety of methods have been proposed for automatically estimating speaking rate from the acoustic signal as a means of adapting recognition algorithms to the speaker's tempo [11] [12].

The speaking rate of each utterance in the diagnostic material was measured using two different metrics. The first, MRATE, derives its estimate of speaking rate from the modulation spectrum of the acoustic signal [12]. Because the modulation spectrum is systematically related to syllable duration [6], MRATE should in principle provide an indirect measure of syllable rate. The second metric used is based directly on the number of syllables spoken per second and is derived from the transcription material.

Figure 16 illustrates the relation between MRATE and word-error rate. Word error does not change very much as a function of MRATE. In many instances the highest error rates are associated with the middle of the MRATE range, while the flanks of the range often exhibit a slightly lower proportion of word errors.

Figure 17 illustrates the relation between word-error rate and syllables per second. In contrast to MRATE, this linguistic metric exhibits a much higher correlation

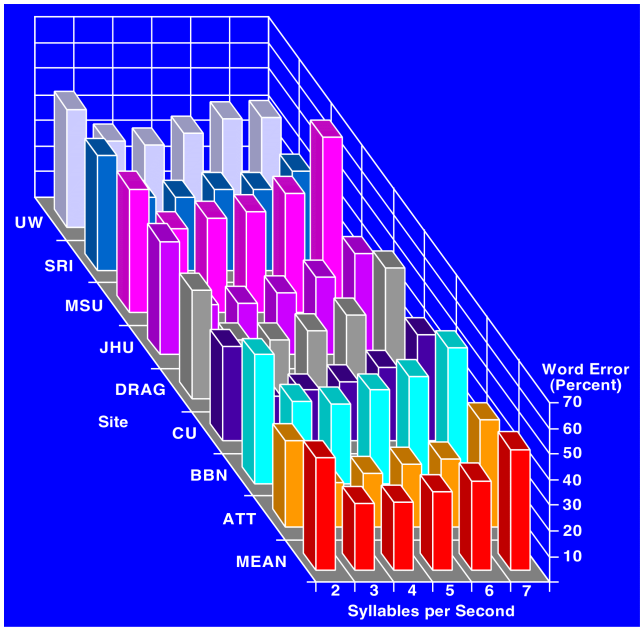

Figure 17: The relationship between word-error rate for each of the eight ASR systems (as well as the mean) and a *linguistic* measure of speaking rate (syllables per second).

between abnormal speech tempo and ASR performance. Utterances slower than 3 syllables/sec or faster than 6 syllables/sec have 50% more word-recognition errors than their counterparts in the core of the normal speaking range. Such data imply that algorithms based on some form of linguistic segmentation related to the syllable are more likely to provide accurate estimates of speaking rate than those based on purely acoustic properties of the speech signal (which may include hesitations, filled pauses and other non-linguistic events that potentially distort the estimate of speaking rate).

12. CONCLUSIONS

The diagnostic evaluation of eight Switchboard-corpus recognition systems suggests that superior recognition performance is closely associated with the ability to accurately classify a word's phonetic constituents. This conclusion is supported by several different forms of analyses, including decision trees and correlations between the magnitude of word and phonetic-segment/articulatory-feature errors.

Additional analyses indicate that syllable structure is also an important factor accounting for recognition performance, as is prosodic stress and speaking rate.

Together, these analyses suggest that future-generation ASR systems should strive for highly accurate phonetic classification and integrate information about syllable structure and prosodic stress into the recognition process.

ACKNOWLEDGEMENTS

The authors wish to thank our colleagues at AT&T, BBN, Cambridge University, Dragon Systems, Johns Hopkins University, Mississippi State University, SRI International and the University of Washington for providing the material upon which the diagnostic evaluation of the Switchboard recognition systems is based. We would also like to express our appreciation to Jeff Good and Leah Hitchcock for prosodically labeling the diagnostic component of the Switchboard corpus, to Joy Hollenback and Rosaria Silipo for assistance with the statistical analyses and data collection, and to George Doddington, Jon Fiscus, Hollis Fitch, Jack Godfrey and Joe Kupin for their helpful comments and advice concerning the analyses described. This research was supported by the U.S. Department of Defense.

REFERENCES

[1] Beckman, M., "Stress and Non-Stress Accent," *Fortis*, Dordrecht, 1986.

[2] Chang, S., Shastri, L. and Greenberg, S., "Automatic phonetic transcription of spontaneous speech (American English)," *Proc. Int. Conf. Spoken. Lang. Proc.*, 2000. (available at http://www.icsi.berkeley.edu/~steveng)

[3] Doddington, G., "Evidence of differences between lexical and actual phonetic realizations," Presentation at the *NIST Speech Transcription Workshop*, College Park, MD, May 18, 2000.

[4] Fosler-Lussier, E., Greenberg, S. and Morgan, N., "Incorporating contextual phonetics into automatic speech recognition." *Proc. XIVth Int. Cong. Phon. Sci.*, 1999. (available at http://www.icsi.berkeley.edu/~steveng)

[5] Godfrey, J. J., Holliman, E. C., and McDaniel, J., "SWITCHBOARD: Telephone speech corpus for research and development," *Proc. IEEE Int. Conf. Acoust. Speech Sig. Proc.*, pp. 517-520, 1992.

[6] Greenberg, S., "Speaking in shorthand - A syllable-centric perspective for understanding pronunciation variation," *Speech Communication*, 29, 159-176, 2000.

[7] Greenberg, S., "The Switchboard Transcription Project," *Research Report #24, 1996 Large Vocabulary Continuous Speech Recognition Summer Research Workshop Technical Report Series. Center for Language and Speech Processing*, Johns Hopkins University, Baltimore, MD, 1997.

[8] Greenberg, S., Chang, S., and Hollenback, J., "An introduction to the diagnostic evaluation of Switchboard-corpus automatic speech recognition systems," *Proc. NIST Speech Transcription Workshop*, 2000. (available at http://www.icsi.berkeley.edu/~steveng)

[9] Lehiste, I., "Surpasegmentals," in *Principles of Experimental Phonetics*, N. Lass, editor. St. Louis: Mosby, 1996.

[10] Martin, A., Pryzbocki, M., Fiscus, J., and Pallet, D., "The 2000 NIST evaluation for recognition of conversational speech over the telephone," *Presentation at the NIST Speech Transcription Workshop*, College Park, MD, May 17, 2000.

[11] Mirghafori, N., Fosler, E., and Morgan, N., "Fast speakers in large vocabulary continuous speech recognition: Analysis and antidotes," *Proc. Eurospeech*, pp. 491-494, 1995. (available at http://www.icsi.berkeley.edu/~nikki)

[12] Morgan, N., Fosler, E., and Mirghafori, N., "Speech recognition using on-line estimation of speaking rate," *Proc. Eurospeech*, pp. 2079-2083, 1997. (available at http://www.icsi.berkeley.edu/~nikki)

[13] Quinlan, J.R., "C4.5: Programs for Machine Learning," *Morgan Kaufmann*, San Mateo, CA, 1993.

[14] Shastri, L. Chang, S. and Greenberg, S., "Syllable detection and segmentation using temporal flow model neural networks," *Proc. XIVth Int. Cong. Phon. Sci.*, pp. 1721-1724, 1999. (available at http://www.icsi.berkeley.edu/~steveng)

[15] Silipo, R. and Greenberg, S., "Prosodic stress revisited: Re-examining the role of fundamental frequency," *Proc. NIST Speech Transcription Workshop*, College Park, MD, 2000. (available at http://www.icsi.berkeley.edu/~steveng)

[16] Silipo, R. and Greenberg, S., "Automatic transcription of prosodic prominence for spontaneous English discourse," *Proc. XIVth Int. Cong. Phon. Sci.*, pp. 2351-2354, 1999. (available at http://www.icsi.berkeley.edu/~steveng)