RESPONSE TIME AS A METRIC FOR COMPARISON OF SPEECH RECOGNITION BY HUMANS AND MACHINES

Anne Cutler and Tony Robinson

MRC Applied Psychology Unit
15 Chaucer Rd.
Cambridge CB2 2EF, U.K.

Dept. of Engineering
University of Cambridge
Cambridge CB2 1PZ, U.K.

ABSTRACT

The performance of automatic speech recognition systems is usually assessed in terms of error rate. Human speech recognition produces few errors, but relative difficulty of processing can be assessed via response time techniques. We report the construction of a measure analogous to response time in a machine recognition system. This measure may be compared directly with human response times. We conducted a trial comparison of this type at the phoneme level, including both tense and lax vowels and a variety of consonant classes. The results suggested similarities between human and machine processing in the case of consonants, but differences in the case of vowels.

1. INTRODUCTION

How can the success of a speech recogniser be evaluated? The obvious way is simply to score a recogniser's output in terms of the number of units - sentences, words, phonemes - which correspond to what was "really" there in the input. This amounts to comparing the recogniser's performance with that of an ideal human listener, who is expected to perform at ceiling and recognise everything correctly. In reality, however, human listeners do not always recognise everything correctly; and even when they do, they find some utterances more difficult to process than others.

An alternative approach to system evaluation, therefore, might be to compare relative difficulty experienced by the machine with relative difficulty experienced by the human listener. Note that this does not address the question of *how* the machine and the human are processing speech, which is chiefly of interest where a machine has been specifically designed to mimic human processing; relative difficulty is still a measure of output success, independent of the internal structure of the recogniser. Relative difficulty ought in principle to depend entirely on recogniser-external factors such as confusability of an input unit with other units in the input repertoire, the intrinsic amount of information in the relevant unit (e.g. its duration), etc. Therefore if a human-machine comparison of relative difficulty were to reveal points at which the machine encountered difficulty but humans did not (or *vice versa*), it might point to ways in which recogniser design could be improved.

To assess relative difficulty for human listeners, it is of course necessary to move human performance off the ceiling. This can easily be done by degrading the input, but in the present case to do so would in effect vitiate the comparison with machine performance since machine and human would no longer be processing the same input. A standard psychological approach to the assessment of processing difficulty is, instead, to measure latency to produce a response of some kind [1]. Response time (RT) is widely used in the study of human speech recognition as a measure of relative processing difficulty at all levels [2] - including the sentence, word and phoneme levels, i.e. the units over which recogniser performance is usually assessed.

We here present a first approach to a comparison of relative processing difficulty via response times of human and machine recognisers. The processing level which we chose to assess is the phoneme. Human RT to detect phonemes is measured by asking listeners to press a response key as soon as they can after being presented with an occurrence of a pre-specified target phoneme; typically the input within which the target is to be detected will be words or sentences. The phoneme detection task has been extensively used as a tool for studying a range of psycholinguistic variables, such as word recognition [3], prosodic processing [4], or the units of prelexical processing [5]; however, it has also produced a considerable amount of data on detection of particular phonemes. It is not the case that for humans any phoneme in any context is equally easy to detect; instead, there is quite a range of human performance, making an informative comparison with a machine analogue of RT a feasible undertaking. In the present study we constructed such a measure for a recognition system, and compared the results it produced for a range of phonemes to human RTs for the same phonemes.

The results are not, however, presented here as an evaluation of the recognition system we used. The purpose of the present study was merely to test the feasibility of comparing human and machine response times; it was, for instance, not possible to conduct the comparison across a single standard input. The contribution of the present report consists in the description of the technique we used and the methods by which we compared the results it produced to the results available from studies of human recognition.

2. THE MACHINE "RTs"

The comparison was conducted using a recogniser with a standard structure: a preprocessor to parameterise frames of speech, an estimator of the probabilities of the class labels for each parameterised frame, and a segmenter and labeller to produce the symbolic sequence of phonemes. Typically a recogniser uses short-term power spectra (or close derivative) as parametric representation, vector quantisation or Gaussian mixtures as probability estimator, and hidden Markov models to produce the most probable phoneme sequence [6]. The approach used here is unusual in the use of recurrent connectionist models for the phoneme probability estimation [7]. This structure is computationally more powerful than the conventional one, and yields a slightly lower error rate [8].

The preprocessor calculates estimates of power, power spectral density, pitch and degree of voicing. Apart from smoothing associated with the pitch frequency, the preprocessor contains no history. The class labels are the 61 phonetic symbols of the TIMIT database. For each frame the *a-posteriori* probability of class occupancy is calculated using the recurrent network. The use of feedback within the net allows past context to be used, and the decision is delayed by four frames (64ms) to allow a limited future context. The *a-posteriori* probability estimator implicitly incorporates the *a-priori* class occurrence probabilities.

Phonemes are modeled as a single state per phoneme Markov model. The maximum likelihood symbol sequence is computed with the Viterbi algorithm [9]. Transition probabilities are obtained by counting from the hand labels and the emission probabilities are provided from the recurrent network.

The process of segmentation and labeling using the Markov model provides the most likely sequence of phonemes for a sentence. Normally this is achieved by making a forward pass over the sentence, and computing for each phoneme the probability of all possible previous phonemes. Identity and likelihood of the most probable previous phoneme are recorded, so that at the end of the sentence it is possible to back track and compute the most likely phoneme sequence. However, in practice there is a time after which any most likely phoneme is independent of future acoustic information. This point is found by tracing the back pointers until they all go through the same phoneme. The minimum time required for a phoneme to become such a point of convergence was the designated RT.

The recogniser processed the TIMIT database, and RTs were calculated as above for all occurrences of the 15 selected phonemes. Error rates were also calculated for each phoneme, as well as the correlation between RT and measured phoneme duration across tokens.

3. THE HUMAN DATA

A substantial body of human RT data from a single subject population was available for 15 phonemes - seven vowels and eight consonants. Of the vowels, six were full vowels: three tense (/a/, /i/, /u/) and three lax (/ɛ/, /ɪ/, and /ʌ/). The seventh vowel was the reduced vowel /ə/. Among the consonants, there were two stops (/p/, /t/), two fricatives (/s/, /v/), two nasals (/m/, /n/) and two semi-vowels (/w/, /j/). Inclusion of a variety of phoneme classes ensured a wide range in the RT distribution. Error (missed detection) data was also available.

The data came from six experiments, four of which have been reported elsewhere [10, 11, 12], while a further two are reported to this meeting [13]. In these experiments human listeners were presented with isolated words (or, in one study, non-words), and were instructed to press a response button as soon as they detected an occurrence of a particular target phoneme. The stimulus materials were blocked such that subjects were listening for only one target at a time; each subject listened for at least four phoneme targets. A total of 171 listeners, all from the Cambridge University community, took part in the experiments, at least 24 in each. /a/ was heard by 147 listeners, /i/ by 96, the four short vowels by 75, and the consonants and /u/ by 24.

The following measures were computed: average RT for each phoneme across listeners; mean error rate (missed responses), ditto; and the correlation between RT and measured phoneme duration for each token in the experiments (token number varied since some phonemes were used in more than one experiment).

4. THE COMPARISON

Our first comparison, of human and machine error rates, showed (unsurprisingly) that error rates were significantly higher for machine than for human performance (t [14] = 8.39, $p < .001$). However, there was a significant positive correlation between the two rates across the 15 phonemes (r [14] = .75 $p < .001$).

These correlations are encouraging since they suggest that the human and the machine results may be tapping similar dimensions of difficulty. However, the correlations may be spuriously produced by differences between (but not within) independent subsets of the data. Therefore we considered the vowel and consonant subsets separately. Separate correlations between the human and machine error rates for vowels versus consonants are shown in Fig. 1; both are at least marginally significant (t [6] = .76, $p < .05$ for vowels, t [7] = .64, $p < .09$ for consonants). It seems that machine errors across the phoneme set are indeed more or less in proportion to human errors.

Table 1 shows the phonemes in order of percentage error by humans and by the machine. The human errors show an interpretable pattern. Among the vowels, lax vowels produce more errors than tense (with the reduced vowel /ə/ producing most errors of all). Exactly this result - fewer errors for tense vowels than for lax - occurs in perceptual confusion studies with human listeners [14]. In the consonants, the greatest proportion of errors occurs on semivowels and the smallest on nasals, with stops and fricatives in between. Again this is similar to the pattern found in confusion data from human listeners [15]. The machine patterns are not as clearly grouped by vowel type or by consonant manner of articulation, but the order is not markedly different from that in the human results.

Table 1. Phonemes in order of mean percentage of errors, from lowest (left) to highest (right), separately for vowels and consonants and for humans *vs.* machine.

Vowels	
Human	i u a ɛ ɪ ʌ ə
Machine	i a ɛ ʌ u ɪ ə
Consonants	
Human	m n t s v p w j
Machine	s n p t w m v j

For RT, a direct comparison of human versus machine RTs is meaningless. However, there was again a significant positive correlation across the phonemes between human and machine performance (r [14] = .53, $p < .05$). Differences between human and machine performance appear, though, when the RT results are broken down into subsets. As Fig. 2 shows, there is a significant correlation between human and machine RTs to consonants (t [7] = .95, $p < .001$), but no relation between human and machine RTs to vowels.

Table 2 spells out the difference. The order of consonant RTs by the machine, is, as the high positive correlation would suggest, very similar indeed to the order produced by the human listeners. The machine orders the vowels quite differently, however (note, for instance, that the machine responds fast to /ə/, which produces the slowest RTs by far from the humans).

We undertook one further analysis to examine the difference between human and machine RTs. The studies of human RTs had consistently found that RT to vowels showed a significant negative correlation with measured phoneme duration: the longer the vowel, the faster the RT [10, 11, 12, 13]. No such systematic relationship appeared however between RT and duration of consonants. Since measured duration was available for all the phoneme tokens used in the human

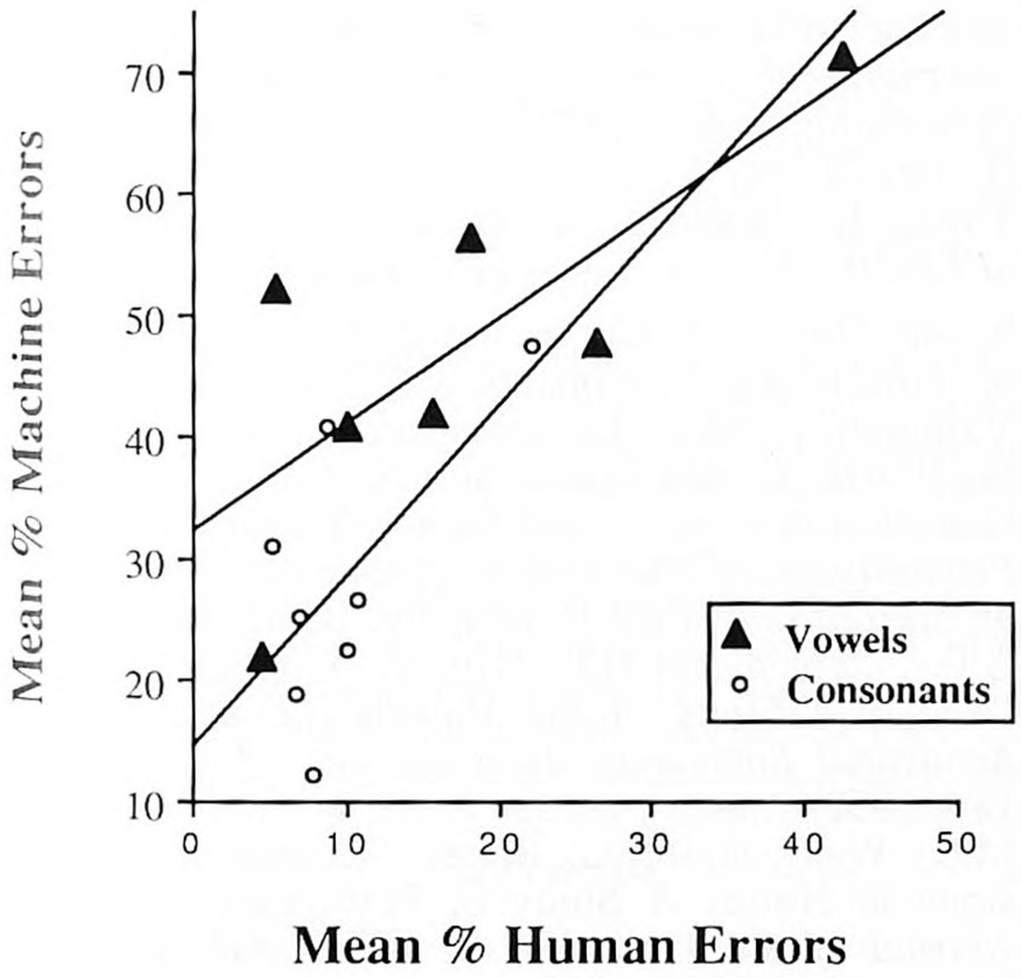

Figure 1. Mean percentage of errors made by human listeners *vs.* the machine for each phoneme, separately for vowels and consonants.

Table 2. Phonemes in order of mean RT, from fastest (left) to slowest (right), separately for vowels and consonants and for human *vs.* machine RTs.

Vowels	
Human	u i a ɛ ʌ ɪ ə
Machine	i ə ɪ ʌ ɛ u a
Consonants	
Human	v n t m p s w j
Machine	v m n p t s w j

experiments, and was also available in the TIMIT labels, we repeated this analysis for each phoneme for both human and machine RTs. The relevant correlation coefficients appear in Table 3. For the human RTs, vowels and consonants produce clearly different results: all vowels show a negative correlation between RT and duration (and all but two of these are statistically significant), while for the consonants, the pattern is unsystematic (and no correlation is significant). The machine RTs also pattern differently for vowels and for consonants. The consonant pattern resembles that found with the human results in that all correlations are close to zero. The vowel pattern produces, on the other hand, a significant correlation for every vowel - but in contrast to the human results, the correlation is positive: the longer the vowel, the slower the RT.

For both human and machine results there was little evidence of a speed-accuracy tradeoff in the vowel-consonant difference: vowels produced longer RTs *and* more errors than consonants. However the relation between mean error rate and mean RT across the phoneme set was closer for the human results (where the two measures were significantly correlated both overall and in vowel and consonant subsets) than for the machine. For humans, added vowel information over time speeds RT and reduces errors; for the machine it slows RT and leaves accuracy unchanged.

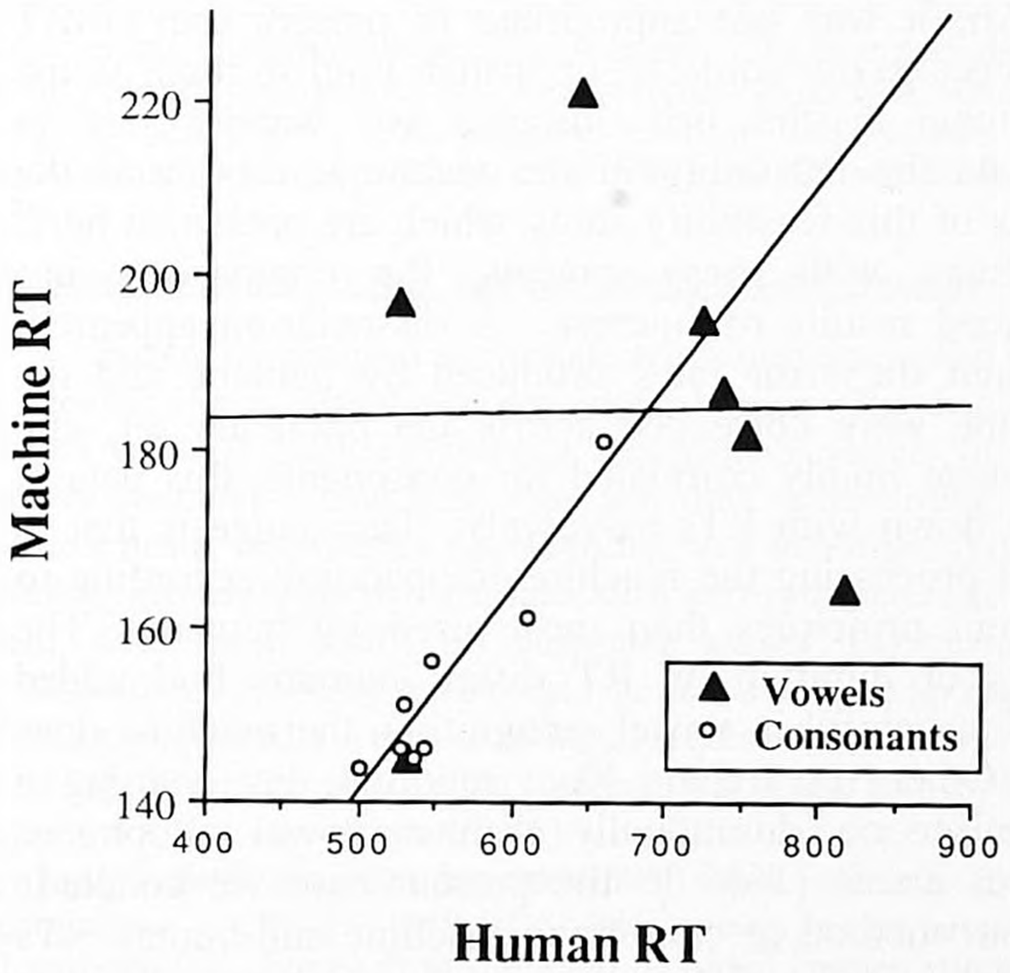

Figure 2. Mean human *vs.* machine RT for each phoneme, separately for vowels and consonants.

Table 3. Correlation coefficients (r) between measured phoneme duration and mean human *vs.* machine RTs.

Vowels	Human	Machine
u	-.33	.30
i	-.22	.11
a	-.26	.47
ɛ	-.43	.64
ʌ	-.20	.61
ɪ	-.15	.42
ə	-.32	.32
Consonants		
m	-.07	.03
n	.07	.14
p	-.22	-.07
t	-.26	.09
v	-.19	-.08
s	.11	.13
w	.15	.17
j	-.31	.07

5. CONCLUSION

We reiterate that this has been merely a feasibility study for a comparison of machine and human RTs. This undertaking arose from a project involving both psychologists studying human speech recognition and engineers constructing a machine recogniser; the processing of phonemes was of interest to both groups, and we sought a measure which would directly compare relative difficulty of phoneme processing for human and machine presented with identical input. The present study is not an objective test of our recogniser's performance, because the input processed by the recogniser and by the human listeners was in fact not identical. Our recogniser was trained on the TIMIT database of American English, while our subject population was trained on (i.e. native in) British English; it was not appropriate to present the TIMIT sentences to our subjects, or British English input to the machine. In the first instance we wanted just to evaluate the feasibility of the technique, and it is the results of this feasibility study which are presented here.

Even with these caveats, the comparison has produced results of interest. A dissociation appeared: although the error rates produced by humans and the machine were correlated across the phoneme set, and RTs were highly correlated for consonants, this pattern broke down with RTs to vowels. This suggests that in vowel processing the machine is operating according to different principles than those used by humans. The effects of duration on RT differ: humans find added duration helpful in vowel recognition, the machine does not. Other researchers have reported that training a recogniser on durationally distinct vowel allophones reduces errors [16]. In the present case we conclude that our method of comparing machine and human RTs has highlighted a difference in processing characteristics for a subset of phonemes, and that such a result in a real test of machine performance could point to aspects in which recogniser performance could be improved.

6. ACKNOWLEDGEMENT

This research was supported by ESPRIT Basic Research Actions, project P3207 "ACTS".

7. REFERENCES

[1] R.G. Pachella. "The Interpretation of Reaction Time in Information-Processing Research". In B.H. Kantowitz (Ed.) *Human Information Processing: Tutorials in Performance and Cognition.* Hillsdale, N.J.: Erlbaum. 1974.

[2] A. Cutler and D. Norris. "Monitoring Sentence Comprehension". In W.E. Cooper and E.C.T. Walker (Eds.) *Sentence Processing.* Hillsdale, N.J.: Erlbaum. 1979.

[3] A. Cutler, J. Mehler, D. Norris and J. Segui. "Phoneme Identification and the Lexicon", *Cognitive Psychology,* vol. 19, pp. 141-177, 1987.

[4] A. Cutler. "Phoneme-Monitoring Reaction Time as a Function of Preceding Intonation Contour", *Perception & Psychophysics,* vol. 20, pp. 55-60, 1976.

[5] D.J. Foss and M.A. Blank. "Identifying the Speech Codes", *Cognitive Psychology,* vol. 12, pp. 1-31, 1980.

[6] L.R. Rabiner. "A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition". *Proceedings of the IEEE*, vol. 77, pp. 257-286, 1989.

[7] T. Robinson and F. Fallside. "A Recurrent Error Propagation Network Speech Recognition System", *Computer Speech & Language*, vol. 5, pp. 259-274, 1991.

[8] T. Robinson. "Several Improvements to a Recurrent Error Propagation Network Phone Recognition System", Technical Report CUED/F-INFENG/TR.82, Cambridge University, 1991.

[9] A.J. Viterbi and J.K. Omura. *Principles of Digital Communication and Coding.* New York: McGraw-Hill. 1979.

[10] A. Cutler, D. Norris and B. van Ooyen. "Vowels as Phoneme Detection Targets". *Proceedings of the International Conference on Spoken Language Processing*, Kobe, vol. 1, pp. 581-584, 1990.

[11] B. van Ooyen, A. Cutler and D. Norris. "Detection Times for Vowels versus Consonants". *EUROSPEECH '91*, vol. 3, pp. 1451-1454, 1991.

[12] B. van Ooyen, A. Cutler and D. Norris. "Detection of Vowels and Consonants with Minimal Acoustic Variation", *Speech Communication*, vol. 11, 1992.

[13] D. Norris, B. van Ooyen and A. Cutler. "Speeded Detection of Vowels and Steady-State Consonants. *Proceedings of the 2nd International Conference on Spoken Language Processing*, Banff, 1992.

[14] G.E. Peterson and H.L. Barney. "Control Methods Used in a Study of the Vowels", *Journal of the Acoustical Society of America*, vol. 24, pp. 175-184, 1952.

[15] M.D. Wang and R.C. Bilger. "Consonant Confusions in Noise: A Study of Perceptual Features", *Journal of the Acoustical Society of America*, vol. 54, pp. 1248-1266, 1973.

[16] L. Deng, M. Lennig and P. Mermelstein. "Use of vowel duration information in a large vocabulary word recognizer". *Journal of the Acoustical Society of America*, vol. 86, pp. 540-548, 1989.