On Speech Recognition : A Statistical Analysis

Author's:

Saleha Rizvi
Danish Nadeem
Haider Jamal Naqvi

Department of Computer Science
Jamia Millia Islamia,
Jamia Nagar, New Delhi-110025

Ranjit Biswas*

Abstract: In speech recognition systems, there is always a possibility of the occurrence of recognition errors due to many factors like speaking style, pronunciation, noise, etc. which are known as human factors. As a consequence there is a necessity of checking the input digits/strings heard /received by the machine so that the error(s), if any, may be detected and rectified. The usual process of spoken input is by string of digits of some suitable length; and thus errors are detected and corrected by feedback after every string of the above length. Naturally time for input is a statistical variate depending upon the string length and the recognition rate where the later differs from machine to machine and is dependent upon the human factors mainly. This paper reveals a statistical analysis of this variate.

Keywords : Recognition error, overhead ratio, positive human factor, negative human factor, overhead time

---

*Address of Correspondence:

Ranjit Biswas,Department of Computer Science,
Jamia Hamdard University, Hamdard Nagar,
New Delhi-62
INDIA.

E-mail : saleha_rizvi@hotmail.com
Phone : 91-11-6829247

1. INTRODUCTION

Speech recognition process consists of some fundamental stages viz. speaking a digit by a human, recognizing the digit by a machine, synthesizing it, speaking the digit by the machine, recognizing the digit by the human ,correcting it if error is there and then repeating the all of the above; else doing the above for the next digit, etc. Thus a factor, which we call by human factor, always exists there due to variations in speaking styles, pronunciation etc. from human to human. As a result checking of digits are to be done for correcting errors, if any. This error is known as recognition error. For instance , if a person asks a telephone operator to obtain a number, the operator repeats every digit (or every string of some fixed and suitable size) in order to be sure that he has heard the telephone number correctly. In case the operator realizes that he has not heard the number correctly, he requests the person to repeat the speaking of the digit (or of the string) . Thus there is a feedback system which ultimately helps in entering digits of the telephone number correctly. During this feedback, the operator or the person who speaks chooses, probably without any statistical or mathematical calculations but conventionally just, his suitable length of the string and perform the job. Similarly when a word is entered into a computer by voice, recognition error may occur due to many human factors like poor pronunciation , noise or environmental constraints. In case of telephone number processing, verification is usually done after every two or three digit string . But in case of entering digits by voice in computer, because of fast processing capability of the machine, long string may be used for a unit of verification or confirmation. But how much long ? Probability of recognizing error increases with the length of the string , and it is also assumed here that errors are to be corrected by repeating the string of digits. Consequently , we are interested in a statistical analysis on the following:

(i) time t to enter a string of length s once

(ii) expected number(N) of strings to be spoken to enter a string correctly.

(iii) expected time (t_s) to enter a string correctly.

(iv) expected time (t_d) to enter a digit correctly.

(v) values of s , if exist , for which t_s and t_d are minimum.

(vi) change of this optimal s with the overhead ratio r .(r is defined below).

This paper follows, mainly , the work of Ainsworth [2]. In [2], Ainsworth put much importance on t_d (defined by his equation number (9) on page 575). He did some useful experimental works and reported some interesting results with graphical representations. In this paper we add to his [2] paper some statistical analysis on some important variates he dealt with , and we finally point out that the mean time per digit (i.e. t_d) with error correction (as defined by (9) in [2]) is not of much importance for discussion ; rather, the time t_s ,which is the expected time to enter a string correctly, is to be studied and analyzed.

2. A STATISTICAL ANALYSIS :-

We suppose that:

t_1 = time required for the human to speak and the machine to recognize a digit.

t_2 = time for the machine to synthesize and speak a digit.

t_0 = time required for the human to change from speaking to listening and the machine to change from recognizing to synthesizing, and vice versa. This is known as "overhead time"[2] **.**

Suppose , feedback is to be given after every string of length s. Time required to speak, recognize and check one string i.e. time required to enter a string is given by

$$t = (t_1+t_2)s + t_0 \quad \text{——————————} \quad (1)$$

and, time required to enter n digits is

$$t(n) = ((t_1+t_2)s+t_0)k+n_0((t_1+t_2)+t_0/s) \quad \text{——————} \quad (2)$$

where $n = k.s + n_0$; $k , s , n_0 (< s)$ being non-negative integers .

Suppose , recognition rate for the machine for each digit is p. Therefore , the probability of existence of at least one error in recognizing string is

$$q = (1 - p^s) .$$

Suppose , X denotes the statistical variate :-

" Number of strings required to be spoken to enter a string correctly".

Clearly , the density function of X is given by

$$f(x) = q^{x-1} . p^s , \quad x = 1,2,3,4,\ldots\ldots$$

Thus, expected number of strings to be spoken to enter a string correctly is given by

$$E(X) = \sum x. f(x)$$

$$= \sum x. q^{x-1}. p^s$$

$$= 1/p^s$$

Therefore the expected time t_s to enter a string correctly is given by

$$t_s = t.E(X) = ((t_1+t_2)s+t_0) . 1/p^s \quad \text{——————————} (3)$$

And the expected time t_d to enter a digit correctly is given by

$$t_d = t_s/s = ((t_1+t_2)s+t_0) . 1(s.p^s) \quad \text{——————————} (4),$$

which is (9) in [2].

In [2] , Ainsworth studied the optimal behavior of t_d with respect to the string length s. The author here has to say that as the feedback is to be done after every string but not after every digit, there is no necessity of observing the optimal behavior of t_d . Instead we are to study the problem of finding out a value of s , if exists ,for which t_s is minimum in order to minimize the total time for entering n digits correctly. Because the minimality of t_d does not imply the minimality of the quantity (s . t_d) . Thus , in

practical fields , the optimal string length as derived by Ainsworth (in (16) of [2]) will not be in fact the optimal one while desiring to enter n digits earliest and correctly.
For studying the optimal behavior of t_s w.r.t s , we solve

$$dt_s / ds = 0 , \quad \text{which gives } s = - r + \log_p e < 0 .$$

In fact , dt_s / ds is always positive. Therefore t_s is a monotonically increasing function of s. Thus the quantity ($s . t_d$) is a monotonically increasing function of s , where t_d is not . If $s_1 = (r^2 / 4 + r / (\log_e p))^{1/2} - r / 2$, (which is (16) in [2]) , then for all $s < s1$ we have $t_s < t_{s1}$.

Thus , although Ainsworth [2] minimized t_d for $s = s_1$, it is not of importance in the case where feedback is done after every string of length s_1 in order to enter all the n digits correctly and earliest.

$$\text{Now , } \quad \min . t_s = (\text{value of } t_s \text{ for } s = 1)$$
$$= (t_0 + t_1 + t_2) / p , \quad \text{from (3).}$$

Therefore, total time required for entering n digits correctly at earliest is

$$= n . (t_0 + t_1 + t_2) / p , \text{ ———————————— (5)}$$

by giving feedback after every digit.
Consequently, in case of auditory feedback the conclusion drawn by Schurick et. al. [5] to give feedback after each word is undoubtedly a good conclusion . In fact , feedback after every string of variable lengths where each string is meaningful to human in some way is good as the meaningful strings, even if of long lengths, increases recognition rate and decreases recognition error . This type of human factors which improves the speech recognition systems may be called by **Positive Human Factors** , whereas the human factors like poor pronunciation, unusual speaking style ,etc . may be called by **Negative Human Factors** . That one string is meaningful is a positive human

factor because a human only realizes its meaning and identifies it as a meaningful word and this capability expedites the whole speech recognition process . Let us study the above conclusion of Schurick et.al.[5] below: -

3. FOR THE CASE WHERE FEEDBACK GIVEN AFTER EVERY MEANINGFUL WORD :

Consider a case of audio speech recognition where feedback is given after every meaningful word. Thus string length varies for each feedback .Consider the entering of a meaningful word of length s. Suppose ,

t_{1s} = time required for the human to speak and the machine to recognize this word.

t_{2s} = time for the machine to synthesize and speak the word.

t_{0s} = time required for the human to change from speaking to listening and the machine to change from recognizing to synthesizing , and vice-versa.

we call it "word overhead time" .

P_s = recognition rate for the machine for the word.

One can see that the following are true :

(i) $t_{1s} < s.t_1$.

(ii) $t_{2s} < s.t_2$. ——————————— (6)

(iii) $t_{0s} < s.t_0$.

(iv) $p_s > p$.

Therefore, time required to enter this word once = $(t_{1s} + t_{2s} + t_{0s})$.

Suppose , Y denotes the variate :-

"Number of times the word spoken to enter it correctly".

The density function of Y is given by

$$g(y) = q_s^{x-1}.p_s,$$

where

$q_s = 1-p_s$, and y = 1,2,3,4,5 ,……. Now,

$$E(Y) = \Sigma\, y.g(y)$$

$$= 1/p_s .$$

Therefore , expected time to enter this word correctly is

$$t_w = (t_{1s}+t_{2s}+t_{0s})/p_s \quad \text{——————————} \quad (7)$$

We see that $t_w < (s.t_1+s.t_2+s.t_0)/p_s$ [using (6)].

$< (t_1+t_2+t_0).s/p$, [using (6)].

Or $t_w <$ time to enter correctly a word of length s by giving feedback after every digit. [Using (5)].

CONCLUSION: In this paper we have studied the recognition error in speech recognition systems and we have made a statistical analysis giving some results**.**

REFERENCES

[1] Ainsworth, W. A ., Audio feedback for error correction in a digit recognition task. Proceedings of the First Europian Conference on Speech Technology, Edinburgh , 2(1987) 65-68.

[2] Ainsworth, W. A ., Optimization of the string length for spoken digit input with error-correction, Int. Jour. Man Machine Studies. 28(1988) 573-581.

[3] Flanagan, J.L., Speech Analysis, Synthesis and Perception, Springler- Verlag (Berlin), 1965.

[4] Peckham, J., Human factors in speech recognition , in G. Bristow Ed., Electronic Speech Recognition , London : Collins (1986).

[5] Schurick, J.M ., Williges, B.H. & Maynard, J.F., User feedback requirements with automatic speech recognition, Ergonomics. 28(1985)1543-1555.

———— X————