# Conditional Probabilities for IBM Voice Browser Recognition of Letters of the Alphabet

TR 29.3421
June 6, 2001

Matthew W. Hartley
James R. Lewis

IBM Voice Systems

West Palm Beach, Florida

## Abstract

The ability to recognize spoken letters of the English alphabet is an important property of speech engines used in telephony applications. Given the acoustic similarity of spoken letters, though, it is a very difficult and potentially error-prone process. This report describes preliminary recognition accuracy data that developers who work with the IBM voice browser can use to design more effective recovery mechanisms for misrecognitions of spoken letters.

## ITIRC Keywords

Voice spelling
Error recovery
Correction of misrecognitions
Speech recognition
Speech browser
Telephony applications

## Contents

## Introduction

In a desktop speech dictation system (such as IBM ViaVoice[1]), the ability of the system to permit voice spelling is important for the product to support hands-free use, but is not critical because the system includes a keyboard on which most users can produce text. Furthermore, users can see the results of the system's interpretation of their spoken letters in real time, and can take steps to recover from recognition errors as they occur.

Telephony applications do not provide this type of support when users need the system to recognize spoken letters for the purpose of spelling names (or other words) that are out-of-vocabulary or hard to recognize. There are no standard mechanisms for spelling with a telephone keypad, and those that exist all have usability problems of one type or another (Lewis, Potosnak, and Magyar, 1997), especially when the keypad is part of the telephone handset. Speech recognition systems that provide *n*-best lists (a list that contains the most likely word or letter for a given spoken input plus the *n* best alternatives -- see Balentine & Morgan, 1999 for more details) require greater resources than systems that do not. For this reason, some systems provide *n*-best lists for spoken input (including letters), but others do not.

When a system does not provide an *n*-best list, an alternative approach is to determine empirically the distribution of misrecognitions among the letters of the alphabet and to use the data from that distribution to guide error recovery schemes. The purpose of this report is to describe preliminary letter misrecognition data for the IBM WebSphere[2] Speech Browser (version 1.0).

[1] IBM and ViaVoice are registered trademarks of International Business Machines Corp.
[2] WebSphere is a trademark or registered trademark of International Business Machines Corp.

## Method

### Participants

The participants in this study were twelve IBM employees. The sample included nine males and three females. The mean age of the sample was 30 with a standard deviation of 4.15 (ranging from 24 to 37). All but two participants spoke with standard American English accents. One male participant was from Thailand and one male participant was from China.

### Materials

Participants used their phones at work to place calls to Cisco[3] 2600 gateway (and a Cisco 2600 gatekeeper), connecting to voice browser (GA version of the IBM Voice Server 1.0) running a VXML program created for the purpose of collecting user speech. The speakers' audio was captured exactly as it came from the gateway/gatekeeper for maximum validity. After capture, the files were edited with Cool Edit 2000[4] to create a separate file for each speaker's pronunciation of each letter of the English alphabet.

### Procedure

The recordings were played into an accuracy-testing program[5] (using the GA version of the IBM Voice Server SDK 1.0 running on an A20 IBM ThinkPad[6]). This procedure was replicated three times to check for outliers or other indications of unstable data. The output of the program provided the information required for estimating misrecognition rates among the spoken letters.

[3] Cisco is a trademark or registered trademark of Cisco Systems, Inc.
[4] Available from Syntrillium Software Corporation.
[5] Developed by Kevin Horowitz.
[6] ThinkPad is a trademark or registered trademark of International Business Machines Corporation.

# Results

## Misrecognitions

Table 1 shows the raw count data, and Table 2 shows those counts converted to rates (correct recognition rates along the diagonal, misrecognition rates in off-diagonal cells). With twelve speakers and three trials per recording, there are 36 opportunities for error for each letter. The number of times that the system returned a silence timeout condition rather than returning a letter appears in the Timeout column. In one case, the recognizer returned the always-active command "Quiet" in place of "Y".

*Table 1. Raw Count Data*

| Said: | A | B | C | D | E | F | G | H | I | J | K | L | M | N | O | P | Q | R | S | T | U | V | W | X | Y | Z | Total |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| *Returned A:* | 6 | | | | | | | | 1 | | | | | | | | | | | | | | | | | | 7 |
| *B:* | | 30 | | 3 | | | | | | | | | | | | | | | | | | | | | | | 33 |
| *C:* | | | 33 | | | | | | | | | | | | | | | | | 3 | | 6 | | | | 3 | 45 |
| *D:* | | 3 | | 23 | 9 | | | | | | | | | | | | | | | | | 3 | | | | | 38 |
| *E:* | | | | | 17 | | | | | | | | | | | | | | | | | 3 | | | | | 20 |
| *F:* | | | | | | 28 | | | | | | | | | | | | | 15 | | | | | | | | 43 |
| *G:* | | | | | | | 33 | | | | | | | | | 6 | | | | 6 | | 3 | | | | | 48 |
| *H:* | | | | | | | | 27 | | | | | | | | | | | | | | | | | | | 27 |
| *I:* | | | | | | | | | 27 | | | | | | | | | | | | | | | | | | 27 |
| *J:* | 3 | | | | | | | | | 30 | | | | | | | | | | | | | | | | | 33 |
| *K:* | 24 | | | | | | | | | | 36 | | | | | | | | | | | | | | | | 60 |
| *L:* | | | | | | | | | | | | 31 | 2 | | | | | | | | | | | | | | 33 |
| *M:* | | | | | | | | | | | | | 31 | 7 | | | | | | | | | | | | | 38 |
| *N:* | | | | | | | | | | | | | | 26 | | | | | | | | | | 2 | | | 28 |
| *O:* | | | | | | | | | | | | 5 | | | 36 | | | 3 | | | 2 | | | | | | 46 |
| *P:* | | | | 1 | 1 | | | | | | | | | | | 24 | | | | | | | | | | | 26 |
| *Q:* | | | | | | | | | | | | | | | | 3 | 36 | | | 3 | 13 | | | | | | 55 |
| *R:* | | | | | | 1 | | | 7 | | | | | | | | | 29 | | | | | | | 2 | | 39 |
| *S:* | | | | | | 6 | | 3 | | 3 | | | | | | | | | 21 | | | | | | | | 33 |
| *T:* | | | | 6 | 5 | | 3 | | | | | | | | | 3 | | | | 24 | | 3 | | | | | 44 |
| *U:* | | | | | | | | | | | | | | | | | | | | | 21 | | | | | | 21 |
| *V:* | | | | | | | | | | | | | | | | | | | | | | 6 | | | | 8 | 14 |
| *W:* | | | | | | | | | | | | | | | | | | | | | | | 33 | | | | 33 |
| *X:* | | | | | | | | | | | | | | | | | | | | | | | | 34 | | | 34 |
| *Y:* | | | | | | | | | | | | | | | | | | | | | | | | | 32 | | 32 |
| *Z:* | | | | | 3 | | | | | | | | | | | | | | | | | 12 | | | | 22 | 37 |
| *Timeout:* | 3 | 3 | 3 | 3 | 1 | 1 | | 6 | 1 | 3 | | | 3 | 3 | | | | 4 | | | | | 3 | | 1 | 3 | 41 |
| *Quiet:* | | | | | | | | | | | | | | | | | | | | | | | | | 1 | | |
| *Total:* | 36 | 36 | 36 | 36 | 36 | 36 | 36 | 36 | 36 | 36 | 36 | 36 | 36 | 36 | 36 | 36 | 36 | 36 | 36 | 36 | 36 | 36 | 36 | 36 | 36 | 36 | 936 |

*Table 2. Rate Data*

| **Said:** | **A** | **B** | **C** | **D** | **E** | **F** | **G** | **H** | **I** | **J** | **K** | **L** | **M** | **N** | **O** | **P** | **Q** | **R** | **S** | **T** | **U** | **V** | **W** | **X** | **Y** | **Z** |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| *Returned A:* | 0.17 | | | | | | | | 0.03 | | | | | | | | | | | | | | | | | |
| *B:* | | 0.83 | | 0.08 | | | | | | | | | | | | | | | | | | | | | | |
| *C:* | | | 0.92 | | | | | | | | | | | | | | | | | 0.08 | | 0.17 | | | | 0.08 |
| *D:* | | 0.08 | | 0.64 | 0.25 | | | | | | | | | | | | | | | | | 0.08 | | | | |
| *E:* | | | | | 0.47 | | | | | | | | | | | | | | | | | 0.08 | | | | |
| *F:* | | | | | | 0.78 | | | | | | | | | | | | | 0.42 | | | | | | | |
| *G:* | | | | | | | 0.92 | | | | | | | | | 0.17 | | | | 0.17 | | 0.08 | | | | |
| *H:* | | | | | | | | 0.75 | | | | | | | | | | | | | | | | | | |
| *I:* | | | | | | | | | 0.75 | | | | | | | | | | | | | | | | | |
| *J:* | 0.08 | | | | | | | | | 0.83 | | | | | | | | | | | | | | | | |
| *K:* | 0.67 | | | | | | | | | | 1.00 | | | | | | | | | | | | | | | |
| *L:* | | | | | | | | | | | | 0.86 | 0.06 | | | | | | | | | | | | | |
| *M:* | | | | | | | | | | | | | 0.86 | 0.19 | | | | | | | | | | | | |
| *N:* | | | | | | | | | | | | | | 0.72 | | | | | | | | | | 0.06 | | |
| *O:* | | | | | | | | | | | | 0.14 | | | 1.00 | | | 0.08 | | | 0.06 | | | | | |
| *P:* | | | | 0.03 | 0.03 | | | | | | | | | | | 0.67 | | | | | | | | | | |
| *Q:* | | | | | | | | | | | | | | | | 0.08 | 1.00 | | | 0.08 | 0.36 | | | | | |
| *R:* | | | | | | 0.03 | | | 0.19 | | | | | | | | | 0.81 | | | | | | | 0.06 | |
| *S:* | | | | | | 0.17 | | 0.08 | | 0.08 | | | | | | | | | 0.58 | | | | | | | |
| *T:* | | | | 0.17 | 0.14 | | 0.08 | | | | | | | | | 0.08 | | | | 0.67 | | 0.08 | | | | |
| *U:* | | | | | | | | | | | | | | | | | | | | | 0.58 | | | | | |
| *V:* | | | | | | | | | | | | | | | | | | | | | | 0.17 | | | | 0.22 |
| *W:* | | | | | | | | | | | | | | | | | | | | | | | 0.92 | | | |
| *X:* | | | | | | | | | | | | | | | | | | | | | | | | 0.94 | | |
| *Y:* | | | | | | | | | | | | | | | | | | | | | | | | | 0.89 | |
| *Z:* | | | | | 0.08 | | | | | | | | | | | | | | | | | 0.33 | | | | 0.61 |
| *Timeout:* | 0.08 | 0.08 | 0.08 | 0.08 | 0.03 | 0.03 | | 0.17 | 0.03 | 0.08 | | | 0.08 | 0.08 | | | | 0.11 | | | | | 0.08 | | 0.03 | 0.08 |

**Conditional Probabilities**

For the data to be useful for the purpose of intelligent error recovery, it is important to compute them as conditional probabilities ($P(A)|A$). The reason for this is that if a user rejects a returned letter or set of letters as being incorrect, the developer (and by extension, the system) can only know the returned letter(s) -- not the spoken letter(s). What the developer needs to know about the returned letter is the probability distribution of all the other letters given that returned letter -- their conditional probabilities -- the ratio of the number of times the system returned a given letter given that the speaker said that letter divided by the total number of times the system returned that letter. This is much more important information than the standard measurement of recognition accuracy (the ratio of number of times a recognizer returns a letter correctly divided by the number of times speakers said that letter).

For example, the standard recognition accuracy of "A" (from the A-A cell of Table 2) was a very low 17%. Often, when a speaker said "A" the system returned "K" (67% of the time). On the other hand, the conditional probability that the speaker had said "A" when the system returned an "A" (from the A-A cell of Table 3) was a fairly high 86%. In other words, when a speaker said "A" the system was likely to misrecognize it, but if the system returned an "A" there was a very good chance that the speaker actually had said "A".

On the other hand, the standard recognition accuracy of "K" was a perfect 100%. Every time a speaker said "K", the system returned "K". However, the conditional probability that a returned "K" occurred when the speaker actually said "K" was a fairly low 60% because the system was likely to produce a "K" when the speaker actually said "A". This means that a voice-spelling application should have fairly low confidence that a returned "K" is evidence of a spoken "K".

Table 3 shows the conditional probabilities computed using the data from the previous tables.

*Table 3. Conditional Probabilities*

| **Said:** | **A** | **B** | **C** | **D** | **E** | **F** | **G** | **H** | **I** | **J** | **K** | **L** | **M** | **N** | **O** | **P** | **Q** | **R** | **S** | **T** | **U** | **V** | **W** | **X** | **Y** | **Z** |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| *Returned A:* | 0.86 | | | | | | | | 0.14 | | | | | | | | | | | | | | | | | |
| *B:* | | 0.91 | | 0.09 | | | | | | | | | | | | | | | | | | | | | | |
| *C:* | | | 0.73 | | | | | | | | | | | | | | | | | 0.07 | | 0.13 | | | | 0.07 |
| *D:* | | 0.08 | | 0.61 | 0.24 | | | | | | | | | | | | | | | | | 0.08 | | | | |
| *E:* | | | | | 0.85 | | | | | | | | | | | | | | | | | 0.15 | | | | |
| *F:* | | | | | | 0.65 | | | | | | | | | | | | | 0.35 | | | | | | | |
| *G:* | | | | | | | 0.69 | | | | | | | | | 0.13 | | | | 0.13 | | 0.06 | | | | |
| *H:* | | | | | | | | 1.00 | | | | | | | | | | | | | | | | | | |
| *I:* | | | | | | | | | 1.00 | | | | | | | | | | | | | | | | | |
| *J:* | 0.09 | | | | | | | | | 0.91 | | | | | | | | | | | | | | | | |
| *K:* | 0.40 | | | | | | | | | | 0.60 | | | | | | | | | | | | | | | |
| *L:* | | | | | | | | | | | | 0.94 | 0.06 | | | | | | | | | | | | | |
| *M:* | | | | | | | | | | | | | 0.82 | 0.18 | | | | | | | | | | | | |
| *N:* | | | | | | | | | | | | | | 0.93 | | | | | | | | | | 0.07 | | |
| *O:* | | | | | | | | | | | | 0.11 | | | 0.78 | | | 0.07 | | | 0.04 | | | | | |
| *P:* | | | | 0.04 | 0.04 | | | | | | | | | | | 0.92 | | | | | | | | | | |
| *Q:* | | | | | | | | | | | | | | | | 0.05 | 0.65 | | | 0.05 | 0.24 | | | | | |
| *R:* | | | | | | 0.03 | | | 0.18 | | | | | | | | | 0.74 | | | | | | | 0.05 | |
| *S:* | | | | | | 0.18 | | 0.09 | | 0.09 | | | | | | | | | 0.64 | | | | | | | |
| *T:* | | | | 0.14 | 0.11 | | 0.07 | | | | | | | | | 0.07 | | | | 0.55 | | 0.07 | | | | |
| *U:* | | | | | | | | | | | | | | | | | | | | | 1.00 | | | | | |
| *V:* | | | | | | | | | | | | | | | | | | | | | | 0.43 | | | | 0.57 |
| *W:* | | | | | | | | | | | | | | | | | | | | | | | 1.00 | | | |
| *X:* | | | | | | | | | | | | | | | | | | | | | | | | 1.00 | | |
| *Y:* | | | | | | | | | | | | | | | | | | | | | | | | | 1.00 | |
| *Z:* | | | | | 0.08 | | | | | | | | | | | | | | | | | 0.32 | | | | 0.59 |
| *Timeout:* | 0.07 | 0.07 | 0.07 | 0.07 | 0.02 | 0.02 | | 0.15 | 0.02 | 0.07 | | | 0.07 | 0.07 | | | | 0.10 | | | | | 0.07 | | 0.02 | 0.07 |

**Most Likely Substitutions**

From the table of conditional probabilities (Table 3), it is possible to develop a list of the most likely substitutions for a given letter. This list appears in Table 4. In Table 4, uppercase letters indicate substitution probabilities that exceeded .10. Lowercase letters indicate substitutions that occurred during the study, but had substitution probabilities less than .10. For example, if a user rejects a returned "A" in a voice-spelling application, then the letter most likely to have actually been spoken is "I". Fourteen of the letters in Table 4 have only one substitution for which the probability of substitution exceeded 10%. Eleven of the letters don't have any substitutes for which the probability of substitution exceeded 10%, and six of those didn't have any substitutions at all. This means that whenever the system returned these letters (H, I, U, W, X, and Y), the developer could have very high confidence that the speaker actually said that letter. Only two letters (T and G) had two substitutes for which the probability of substitution exceeded 10%.

*Table 4. Most Likely Substitutions*

| **Returned** | | **Likely Substituted For** | | | | |
|---|---|---|---|---|---|---|
| *A* | | I | | | | |
| *B* | | d | | | | |
| *C* | | V | t | z | | |
| *D* | | E | b | v | | |
| *E* | | V | | | | |
| *F* | | S | | | | |
| *G* | | P | T | v | | |
| *H* | | | | | | |
| *I* | | | | | | |
| *J* | | a | | | | |
| *K* | | A | | | | |
| *L* | | m | | | | |
| *M* | | N | | | | |
| *N* | | x | | | | |
| *O* | | L | r | u | | |
| *P* | | d | e | | | |
| *Q* | | U | p | t | | |
| *R* | | I | f | y | | |
| *S* | | F | h | j | | |
| *T* | | D | E | g | p | v |
| *U* | | | | | | |
| V | | Z | | | | |
| W | | | | | | |
| X | | | | | | |
| Y | | | | | | |
| *Z* | | V | | | | |

## Discussion

The data presented in this report are preliminary in the sense that they came from a fairly small sample size. The consequences of the small sample size are that the smaller estimated probabilities might not be perfectly accurate. The larger estimated probabilities, however, are likely to remain stable given an increase in the sample size, and it is these larger probabilities that would play the greatest role in error recovery schemes.

It is also important to note that the conditional probabilities and most likely substitutions should not be the only data that developers bring to bear on the problem of developing their error recovery schemes. If users spell English words, then it would be possible to use published tables of unigram and digram frequencies (Card, Moran, & Newell, 1983) to supplement the conditional probabilities presented in this report. If users "spell" strings that are not English words (for example, codes using alphabetic characters, such as part numbers or membership numbers), then developers should use whatever they know about the characteristics of these strings to guide efficient error recovery.

Despite any shortcomings, these data should be useful to developers who design error recovery schemes for voice spelling tasks that are part of telephony applications.

## References

Balentine, B., & Morgan, D. P. (1999). *How to build a speech recognition application: A style guide for telephony dialogues* . San Ramon, CA: Enterprise Integration Group.

Card, S. K., Moran, T. P., and Newell, A. (1983). *The psychology of human-computer interaction*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Lewis, J. R., Potosnak, K. M., and Magyar, R. (1997). Keys and keyboards. In M. Helander, T. K. Landauer, and P. V. Prabhu (Eds.), *Handbook of Human-Computer Interaction* (pp. 1285-1315). Amsterdam: North-Holland.