The Machine as an Addressee: When Paralinguistics Fails

M. Ohala
San Jose, USA

1. Introduction

Speech normally encodes both linguistic and non-linguistic information. The latter conveys, among other things, the following:

a. The speaker's state, e.g., relaxed, angry, interested, bored, nervous, sad, etc.,
b. The speaker's attitude towards the addressee, e.g., formal, informal, cooperative, aggressive, condescending, etc.,
c. The speaker's attitude toward the content or referent of the message, or, more generally, the way the speaker is reacting to the informational context in which the speech is uttered, especially whether it represents knowledge shared with the addressee or not.

Although such 'paralinguistic' signals are always present in speech, and humans have learned to adapt to them, there is one very new use of speech where such non-linguistic features impair communication, namely when the listener is a machine.

Machines which will recognize spoken commands are now commercially available and are being used in a number of applications: mail sorting, assembly lines, control of wheelchairs by quadraplegics, etc. These automatic speech recognition (ASR) devices are generally speaker-dependent and work on words spoken in isolation. To use one, a given speaker must train the device by first giving it samples of his/her pronunciation of all the words it will be required to recognize (typically $\leqslant 200$). The acoustic pattern of these words (called 'templates') are stored in the device's memory; the incoming 'unknown' word is then compared with all the stored templates and is identified as the word corresponding to the template which produced the closest match. (If no template produces a sufficiently close match the device may prompt the speaker to repeat the word.) Once a speaker gets used to the device, accuracy rates of 97% or better are not uncommon. But the problem is that some users take a long time to 'get used' to such devices (2 to 6 weeks) due to a high degree of initial variability in pronunciation (as Doddington and Schalk (1981) remark: 'Speech recognizers commercially available today are effective only within narrow limits. They have relatively small vocabularies and frequently confuse words. Users must develop the skill to talk to the

recognizer, and the machine's performance often varies widely from speaker to speaker'. p. 26). There would be considerable practical advantage if the source(s) of within-speaker variation could be identified so that strategies to control it/them could be developed. A number of determinants of within-speaker variation in language use have already been identified and described in some detail. Conceivably, at least some of the causes of speaker variability which plagues ASR are one or more of these previously-identified factors.

Among the things that need to be considered in trying to find the causes of speaker variability are the following:

Does the speaker cast himself/herself in some well-defined social role vis-a-vis the ASR device, e.g., as a superior, as an adult speaking to a child, an owner to a dog, an English teacher to a foreigner? Does the perception of role persist or, worse, does it change depending on the type of feedback (or lack of it) received from the device? Very possibly the user finds him/herself in a totally new 'social' situation and discovers that the old and familiar sociolinguistic roles do not apply. It does not pay, however, to be alternately helpful, exasperated, condescending, etc., to an ASR device. To the extent that the speaker 'tries out' various socially-dictated modes of speaking, the ASR device is more likely to fail.

A more subtle source of variation, mentioned in (c), is the speaker's presumption of shared knowledge with the addressee. To oversimplify, whatever the speaker thinks the listener knows, or *should* know, can be weakly articulated. Conversely, whatever represents new information must be pronounced carefully, or at the speaker's option emphatically. Both weakly-articulated and emphatic pronunciation may differ from context-neutral pronunciation, thus creating problems for an ASR device.

2. Experiment

I hypothesized that a significant part of within-speaker variation stems from a speaker using the familiar emotional and attitudinal qualifiers which, though appropriate when communicating with other humans, are inappropriate when speaking to machines. Thus, the more 'emotional' the speaker becomes the more he will vary the way he speaks in order to express that emotion, and such pronunciation variability will lead to degradation of ASR performance.

3. Experimental design

To test this I observed and recorded 20 subjects' interaction with an ASR device under circumstances where their emotional arousal could be controlled. Subjects were randomly assigned to 4 groups of 5 each in an experimental design whereby two binary factors were varied independently, high vs. low subject involvement in (or anxiety towards) the task, and high vs. low confusability of the vocabulary used (such that error rate would be high vs.

low). Involvement in the task varied by paying half the subjects by the number of words recognized correctly and paying the other half by the hour. The vocabularies, one with many phonetically similar words and one with highly dissimilar words provided variation in the inherent difficulty of the recognition task. The assignment of the 4 groups according to these experimental variables is shown in Table I. It was hypothesized that subjects who were paid by the word and assigned the confusable vocabulary would have the highest error rate.

Table I. Experimental design.

Variable 1 Subject interest	No involvement (paid by the hour)	High involvement (paid by the word)
Variable 2 Complexity of voc.		
Low error rate (distinct voc.)	Group I	Group II
High error rate (confusable voc.)	Group III	Group IV

With one exception all subjects were students at the University of California, Berkeley, from various disciplines. Each subject first trained the ASR device by pronouncing a single time each of the 30 words of the vocabulary assigned to him/her. These samples constituted the stored templates. Then, with the computer prompting them via printed words (randomized and in blocks of 30) on the CRT of the terminal, they repeated the words for a total of 450 trials in one session and 450 trials in a second session on another day. In all, thus, each subject had 900 trials. During the recognition session there was an interval of 2 seconds between responses and the next prompt. After 90 trials subjects were given short breaks. All sessions were audiotape recorded for later acoustic analysis. A record of correct/incorrect recognition was automatically maintained by the computer. There was a constant threshold for 'rejection', i.e., when a noise or utterance was judged to be so dissimilar as to be unlike any of the stored templates. Rejections were not counted as errors. (For details on the algorithm used in the recognition see Murvelt, Lowy and Brodersen (1981).

4. Results

Figure 1 shows the average error rate by blocks of 750 words for all 4 groups. As would be expected subjects using the distinct vocabulary (Groups I and II) made fewer errors than those with the confusable one (Groups III and IV). Counter to expectation, those paid by the word (Groups II and IV) did

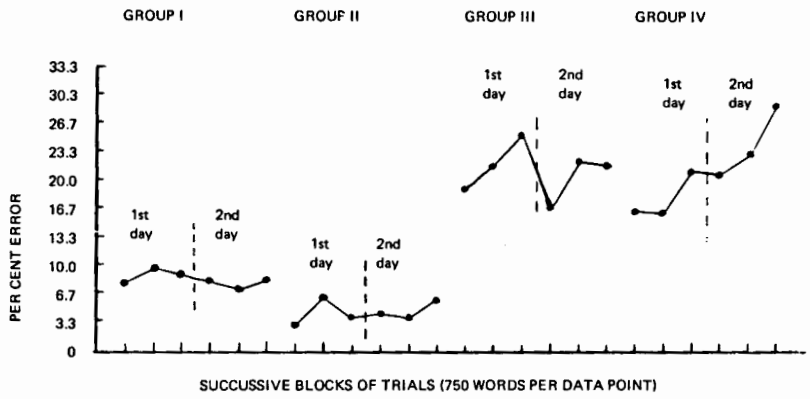

Figure 1. Percent error for the 4 experimental groups plotted as a function of successive blocks of 750 words (see text for further details).

not have more errors than those paid by the hour, however the effect of subject's anxiety on performance is reflected in another interesting way. Groups III and IV (in contrast to the other two groups) give evidence of experiencing an *increasing* error rate on successive blocks of trials. This is unusual because speakers normally adapt to the constraints of these tasks, i.e., manage to lower the error rate. Evidently errors beyond a certain level trigger an emotional reaction in speakers which in turn causes them to express this emotion in their speech thus leading to more errors.

5. Conclusions

The results of this study support the hypothesis that one source of error in ASR is the variation in speakers' pronunciation which encode their changing emotional state. Further studies are underway to identify the precise acoustic features which manifest these emotions.

Acknowledgments

This research was funded by an Affirmative Action Faculty Development Program grant awarded by the Chancellor's Office of the California State University system. The research was conducted using the experimental ASR facility at the Electrical Engineering and Computer Science Department at the University of California, Berkeley. I am extremely grateful to R.W. Brodersen for making this facility available to me and to M. Murveit for his time and guidance in the use of the ASR device and also for writing the computer program. I am also grateful to Margit Peet for her help in conducting the sessions with the subjects and collecting the data, and to John Ohala for comments on the experimental design.

The vocabulary was designed by John Ohala. The distinct vocabulary included words such as *lunch, claud, point, lake* and the confusable vocabulary included words such as *bad, dad, bead, deed.*

References

Doddington, G.R. and Schalk, T.B. (1981). Speech recognition: turning theory to practice. *IEEE Spectrum* **18 (9)**: 26-32.

Murveit, H., Lowy, M., and Brodersen, R.W. (1981). An architecture of an MOS-LSI speech recognition system using dynamic programming. *JASA* **69**: 42.