A MODEL OF AUDITORY-VISUAL SPEECH PERCEPTION

P.J. Blamey and G.M. Clark

Department of Otolaryngology
University of Melbourne

ABSTRACT - A mathematical model relating the probabilities of correctly recognizing speech features, phonemes, and words was tested using data from the clinical trial of a multiple-channel cochlear implant. A monosyllabic word test was presented to the patients in the conditions hearing alone (H), lipreading alone (L), and hearing plus lipreading (HL). The model described the data quite well in each condition. The model was extended to predict the HL scores from the feature recognition probabilities in the H and L conditions. The model may be useful for the evaluation of automatic speech recognition devices as well as hearing impaired people.

INTRODUCTION

A recent publication (Boothroyd, 1985) has shown that the probabilities of correctly recognizing features, phonemes and words may be related by fairly simple mathematical expressions. If the probabilities are measured in a single experiment using monosyllabic words with a consonant-vowel-consonant (CVC) structure, the relations are

$$p = f^l \qquad (1)$$

$$w = p^j \qquad (2)$$

where f, p, and w are the probabilities of correctly recognizing features, phonemes, and words. l and j are constants. l may be thought of as the number of statistically independent features that must be recognized to correctly identify a phoneme. j may be thought of as the number of statistically independent phonemes that must be recognized to identify a word. Since there are three phonemes in a CVC word, we might expect j to be three. However, not all combinations of consonants and vowels are valid words and this lexical constraint produces a value less than three. Similarly, the value of l is affected by phonetic constraints. It is conceivable that the details of the perception process itself may also play a role in determining the values of l and j. They may depend on which features are available from the input signal and the access to lexical and phonetic information. It is of interest to see whether the model is applicable to lipreading as well as hearing, and whether the l and j factors are the same in the two cases.

In this study, the probabilities of correctly recognizing features, phonemes, and words were measured in three conditions: hearing alone (H), lipreading alone (L), and hearing plus lipreading (HL). These probabilities were analysed to see whether equations 1 and 2 adequately described the data. The values of l and j were determined for each condition and tested for equality.

In conjunction with equations 1 and 2, an equation relating feature recognition in the HL condition to the corresponding probabilities in the H and L conditions would allow the prediction of phoneme and word recognition

probabilities in any condition. One such equation may be postulated on the assumption that the auditory and visual inputs provide statistically independent sources of information contributing to feature recognition. This leads to the equation

$$(1-f_{HL}) = (1-f_H)(1-f_L) \div (1-f_C) \qquad (3)$$

where f_{HL}, f_H and f_L are the probabilities of correctly recognizing features in the HL, H and L conditions, and f_C is the probability of correctly recognizing a feature by chance. The value of f_C was estimated and the predicted values of f_{HL} were compared with the observed values.

METHOD

A clinical trial of a multiple-channel cochlear implant (Dowell et al, 1985) furnished the raw data for this study. The patients were all postlingually deafened and so had developed auditory skills through normal hearing. After becoming deaf, they had all relied on lipreading for everyday communication and so were practised lipreaders. Among the patients there was a broad range of test scores, which was necessary for the precise measurement of l and j. The auditory input for the patients came from electrical stimulation within the cochlea by the implanted device.

The patients were tested using videotaped lists of 50 CVC words (Peterson and Lehiste, 1962; Tillman and Carhart, 1966), recorded by an adult Australian speaker. Preoperative results were available for the L condition and postoperative results in all three conditions. Different numbers of tests were carried out with individual patients in different conditions and all available data was included in the analysis below. The probabilities of correctly recognizing features in the three conditions (f_H, f_L, and f_{HL}) were calculated from the responses. The features used were voicing, manner and place of articulation for the consonants and duration, F1 frequency, and F2 frequency for the vowels. The vowels were divided into high, low, and intermediate groups on the basis of formant frequencies from a study of Australian male speakers by Bernard (1970). Probabilities for phoneme recognition (p_H, p_L, and p_{HL}) and for word recognition (w_H, w_L, w_{HL}) were also determined from the responses.

RESULTS

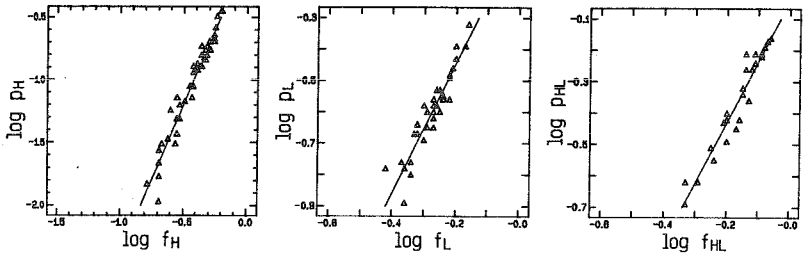

Figure 1. Phoneme versus feature recognition probabilities in the H, L and HL conditions.

Figure 1 shows graphs of log(p) against log(f) for each condition. Figure 2 shows log(w) against log(p). Equations 1 and 2 require that the data lie along straight lines in the graphs as plotted. The slopes of the lines are

estimates of l and j. Each line should pass through the origin where f = p = w = 1. Least squares regression lines were fitted to the data and these are shown in Figures 1 and 2.

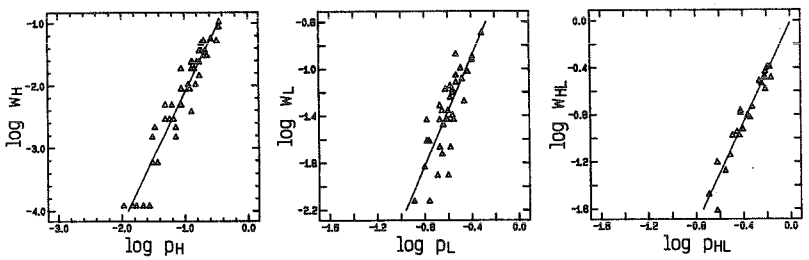

Figure 2. Word versus phoneme recognition probabilities in the H, L and HL conditions.

The l and j values derived from the regression lines are shown in Table 1 together with the intercepts and the standard deviations of these parameters. Different numbers of test lists were available in the different conditions as shown in Table 1. The correlation coefficients between log(p) and log(f) and between log(w) and log(p) indicate that the linear regressions account for a high proportion of the variance in each set of data. t-tests show that none of the intercepts is significantly different from zero so that the data are consistent with straight lines passing through the origin. t-tests of the l and j values show a significant difference between the l values for H and L ($t=3.52$, $df=73$, $p<0.001$), but no differences in the j values ($t=1.06$, $df=73$, not significant).

Table 1. Regression parameters and correlation coefficients.

Condition	Slope	Intercept	n	correlation coefficient
log(p) vs log(f)				
H	2.44 ± 0.09	0.02 ± 0.04	42	0.97
L	2.04 ± 0.14	-0.03 ± 0.04	35	0.93
HL	1.97 ± 0.12	-0.03 ± 0.02	24	0.96
log(w) vs log(p)				
H	2.10 ± 0.10	0.00 ± 0.11	42	0.95
L	2.31 ± 0.27	0.02 ± 0.17	35	0.82
HL	2.14 ± 0.13	-0.02 ± 0.13	24	0.96

The value of f_c was determined by dividing the test lists into pairs and using one as the stimulus list and the other as the response list. The proportion of features that were the same in the two lists was used as an estimate of f_c. The value obtained was 0.38 ± 0.01. Figure 3 shows the measured values of f_{HL} against the values predicted from equation 3. Only data from patients for whom the three conditions were tested on the same day was included in this analysis. This was done to reduce extraneous

variations such as learning effects. A t-test of the differences between the observed and predicted values showed a marginally significant difference equal to 0.014, t=2.06, df=15, p<.10. The correlation coefficient between observed and predicted values was highly significant (r=0.78, t=4.59, df=14, p<0.001).

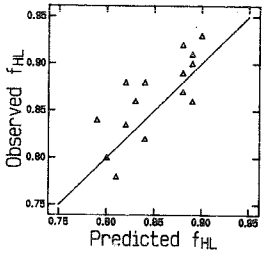

Figure 3. Observed feature recognition probabilities in the HL condition versus values predicted from equation 3.

DISCUSSION

Equations 1 and 2 provided quite an accurate description of the data. The j values appeared to be independent of condition. Boothroyd and Nittrouer (1984) found values of j between 2.05 and 2.54 for normally hearing subjects listening to monosyllabic words in spectrally shaped noise with a variety of signal to noise ratios. Boothroyd (1985) obtained a value of 2.3 for j from data derived from a study of the perception by normally hearing listeners of the speech of deaf speakers. These values are close to those found here. Boothroyd and Nittrouer found that the j factors for lists of frequently occurring words were lower than those for lists of less frequently used words. This effect is not relevant to the present study since the lists of words were approximately balanced for word frequency. In general, the values of j and l may be different for different word lists.

The difference between the l values for the H and L conditions may have arisen from the different speech features used in hearing and lipreading. The values obtained depended on the features chosen for the analysis: voicing, manner, place, F1, F2, and duration. This choice was made for convenience and may not be the most appropriate one for this type of model. The features divide the phonemes into different numbers of groups (for example, 2 groups for voicing, 3 groups for formant frequencies, 5 groups for manner). The features are also perceived with varying degrees of difficulty (for example, voicing is very difficult to perceive through lipreading while some aspects of place are perceived easily). The model ignores these facts and assumes that all phonemes and all features have equal recognition probabilities. Despite these apparent flaws, the model produces an empirical description of the data that is remarkably good. Boothroyd (1985) measured an l factor for binary contrasts of features in phonemes, but this technique is not adaptable to measurement using open set CVC words. Because it was based on different features and different measures, Boothroyd's value of 4.3 for l is not comparable to those found in this study.

Figure 3 shows that equation 3 provided a reasonable empirical basis for the calculation of f_{HL} from f_H and f_L. The result was a slight underestimate of f_{HL}. A detailed discussion of the combination of auditory and

visual information is included in a separate study (Blamey, 1986).

Equations 1, 2, and 3 may now be applied to predict the phoneme and word scores in any of the three conditions, starting from values of f_H and f_L. Figure 4 shows the values of w_{HL} calculated from equations 1, 2 and 3 assuming particular values of f_H and f_L. Each curve represents a different value of f_L. As an example, consider the case of a cochlear implant patient who is an average lipreader with f_L equal to 0.75. Equations 1 and 2 predict w_L=0.26. If the implant increases f_H from the chance value of 0.38 to 0.68, the average value observed in this study, then the model predicts w_{HL}=0.56. This is a very worthwhile improvement of over 100% in the recognition of words.

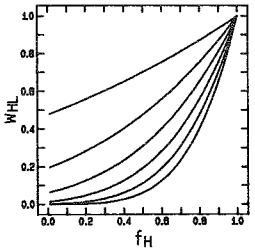

Figure 4. Word recognition probabilities in the HL condition calculated from feature recognition probabilities in the H and L conditions. f_l is 0.4, 0.5, 0.6, 0.7, 0.8, and 0.9 from the lowest to the highest curve respectively.

The model is expected to be useful for predicting the performance of prospective implant patients, for investigating differences between individual patients, and for comparing the results of different types of speech tests. A model of this type may also be useful for the evaluation of automatic speech recognition devices since the j and l factors are measures of the effectiveness with which the lexical and phonetic constraints are used by the device. If alternative input systems are available, they may take the place of the H and L conditions above, and the result of a combined system can be predicted.

ACKNOWLEDGEMENTS

Richard Dowell and Alison Brown carried out the testing of patients for the clinical trial and very kindly made the data available for analysis. Arthur Boothroyd generously discussed his mathematical model at length during his recent visit to Australia. Financial support was given by the National Health and Medical Research Council of Australia.

REFERENCES

BERNARD, J.R.L. (1970) "Toward the Specification of Australian English", Z. Phonetik 23, 113-128.

BLAMEY, P.J. (1986) "Combining Auditory and Visual Information in Speech Perception", in preparation.

BOOTHROYD, A. (1985) "Evaluation of Speech Production of the Hearing Impaired: Some Benefits of Forced-Choice Testing", J. Speech Hear. Res.

28, 185-196.

BOOTHROYD, A. & NITTROUER, S. (1984) "Quantification of Lexical Redundancy Effects in Word Recognition" presented at the Spring 1984 meeting of the Acoustical Society of America, Norfolk, Va.

DOWELL, R.C., MARTIN, L.F.A., CLARK, G.M. & BROWN, A.M. (1985) "Results of a Preliminary Clinical Trial on a Multiple Channel Cochlear Prosthesis", Ann. Otol. Rhinol. Laryngol. 94, 244-250.

PETERSON, G.E. & LEHISTE, I. (1962) "Revised CNC Lists for Auditory Tests", J. Speech Hear. Dis. 27, 62-70.

TILLMAN, T. & CARHART, R. (1966) "An Expanded Test for Speech Discrimination Utilizing CNC Monosyllabic Words", USAF School Aerosp. Med. Rep. SAM-TR-66-55.