COMBINING TACTILE, AUDITORY AND VISUAL INFORMATION FOR SPEECH PERCEPTION

P.J. Blamey and G.M. Clark

Department of Otolaryngology
University of Melbourne

ABSTRACT - Four normally hearing subjects were trained and tested with all combinations of a highly degraded auditory input, a visual input via lipreading, and a tactile input using a multichannel electrotactile speech processor. When the visual input was added to any combination of other inputs, a significant improvement occurred for every test. Similarly, the auditory input produced a significant improvement for all tests except closed-set vowel recognition. The tactile input produced scores that were significantly greater than chance in isolation, but combined less effectively with the other modalities. The less effective combination might be due to lack of training with the tactile input, or to more fundamental limitations in the processing of multimodal stimuli.

INTRODUCTION

The purpose of this study was to investigate combinations of auditory (A), visual (V), and tactile (T) modalities for speech recognition. These modalities have been studied individually and in the combinations AV and VT. There are two further combinations, AT and AVT that are just beginning to be studied (Eilers *et al*, 1988). The clinical situations in which these combinations would be relevant are the cases of severely and profoundly hearing-impaired people who gain some benefit from conventional hearing aids, but not enough to achieve a high level of comprehension. The present investigation was designed as an initial evaluation of the usefulness of tactile information in the AT and ATV conditions, and included equivalent investigations of the A, V, T, AV and VT conditions for comparative purposes. These combinations of sensory inputs have seldom been studied in controlled circumstances with the same set of subjects. The specific questions addressed by the study are whether the combined modalities present more information than the individual modalities, and whether the A, V, and T information combine equally effectively.

The tactile modality differs from the others because speech information is not normally available in a tangible form. An electrotactile multichannel speech processor, the "Tickle Talker" (Blamey & Clark, 1987), was used in the present investigation. Another problem is the availability of experienced users of the tactile device since it is obvious that any person will need an extensive period of training before the newly presented tactile information can become associated with the meaningful perception of speech. This contrasts strongly with the considerable experience most persons have with auditory and visual speech perception. In the present study, four normally hearing listeners who had been trained with the tactile device in a previous study participated as subjects. This situation was far from ideal because of the limited experience of the subjects.

METHODS

Subjects and Training

Four normally hearing subjects took part in this experiment. Each was a female tertiary-level student who was paid for her participation. Their ages ranged from 20 to 27. Each subject reported that she had normal vision, but no formal tests were carried out. Each subject had electrotactile thresholds and comfortable stimulation levels that fell within the normal range. All four had previously been trained with the Tickle Talker over a six-month period, using lipreading but no auditory signal (Cowan *et al*, 1988). In the earlier study, each subject was trained for a total of 70 hours, using the speech tracking procedure of de Filippo and Scott (1978) in the V and VT conditions and closed sets of nonsense syllables and words in V, T and VT conditions. At the conclusion of this six-month period, the subjects showed significant differences between the V and VT conditions on open-set word and sentence recognition tests, on closed-set vowel and consonant tests and also in speech tracking rates. Scores for recognition of closed sets of vowels and consonants were also well above chance in the T condition. A gap of two months occurred between the end of the earlier study and the start of the present one. The subjects did not use the Tickle Talker at all during this time. No specific training was given to the subjects for the present study, but some improvement in scores was observed for those tests that were repeated.

Evaluation Methods and Materials

The four subjects were tested in sessions lasting one hour or two hours with a short break in the middle. Each subject attended 15 to 20 sessions in a two-month period. In the majority of the sessions, the subjects were tested using different sensory modalities with 10 minutes of speech tracking, followed by one or two closed-set recognition tasks. The modalities A, V, T, AV, AT, VT and AVT were tested in rotation, in a different order for each subject. Speech tracking was not done in the T condition because of the difficulty of this task.

Three different closed-set tasks were used: vowel recognition, using the words "hid, head, had, hud, hod, hood, heed, heard, hard, who'd, hoard"; consonant recognition, using the consonants /p,b,m,f,v,s,z,n,g,k,d,t/ in an /a/-consonant-/a/ context; and a set of twelve words containing monosyllables, trochees, spondees and polysyllabic words (MTSP): "fish, ball, shoe, table, pencil, water, airplane, toothbrush, popcorn, elephant, Santa Claus, butterfly" proposed by Erber (1982). Each closed-set task consisted of a block containing four of each stimulus in a randomized order. Results were obtained for three blocks of vowels or consonants in each condition. Each subject scored 100% for the MTSP test in the V condition, so two blocks of results were collected for the A, T and AT conditions, omitting those that included a visual component. The tests were presented with live voice and feedback was given after each item. The speaker for all of the above tests was an Australian male who was previously unknown to the subjects and had not been involved in the training or testing for the earlier study. The speaker was aware of the condition being tested. Results were obtained for six ten-minute sessions of speech tracking in each condition except T.

In the final sessions, the subjects were tested with the open-set Bench, Kowal, and Bamford (BKB) sentence test (Bench & Bamford, 1979), and the open-set Consonant-Nucleus-Consonant (CNC) word test (Peterson & Lehiste, 1962) in each condition. These tests were recorded on videotape using another Australian male speaker. A different test list was used for each condition, and the order of testing the conditions was balanced across the four subjects.

Input Signals

A degraded auditory input was provided to the subjects who were seated in a sound attenuating booth and could not hear the direct signal from the speaker's voice. The speech signal was filtered with a digital elliptic filter with 7 poles and 6 zeroes and a cut-off frequency of 400 Hz. The filtered signal was then amplified again to 80 dBA peak level and mixed with white noise at a level of 70 dBA. The signal was presented binaurally to the subjects through headphones. The white noise (without the filtered speech) was presented in the V, T, and VT conditions to mask the direct voice signal which was reduced by the sound attenuating booth. The measured attenuation was 45 dBA. In the A, AT, AV, and AVT conditions, the white noise also had the effect of masking quiet sections of the filtered speech signal, and high-frequency components that were not completely removed in the filtering process. The acoustic signal was chosen to provide a very crude simulation of a severe hearing loss.

The visual signal was provided via a double glazed window in the sound attenuating room for the live voice testing. The speaker's face was well lit by lamps from both sides of the face and lighting was turned off on the listener's side of the window to avoid reflections in the glass. The total distance between speaker and listener was approximately 1 m. In the recorded tests, the visual signal was presented with a 48 cm color television monitor at a distance of about 1.5 m. The speaker's head was shown completely and occupied about 90% of the vertical extent of the screen.

The tactile signal was provided via the Tickle Talker, a multiple-channel electrotactile speech processor which produced estimates of the fundamental frequency, EF0, the second formant frequency, EF2, and the amplitude, EA, of the speech signal. The value of EF0 was scaled linearly so that a frequency of 250 Hz produced a stimulation pulse rate of 150 pps. In the case of unvoiced sounds, this circuit produced a series of pulses with random time intervals between them. The EF2 range was divided into eight regions, corresponding to the eight electrodes worn by the subjects. There was one electrode on each side of each finger (excluding the thumb) of the left hand, and a common electrode on the left wrist. Each electrical pulse (at the scaled EF0 rate) was applied between the wrist electrode and one of the finger electrodes chosen according to the value of EF2 at the time the pulse was applied. The frequency boundaries between the electrodes were 900, 1125, 1350, 1575, 1800, 2400 and 3300 Hz. Although the second formant does not usually extend as high as 3300 Hz, the output of the EF2 circuit could exceed this value for sounds such as /s/ and /z/ which include intense high frequency components. These components are not second formant resonances, but still provide useful information to the subjects. The amplitude estimate, EA, controlled the duration of the 1.5 mA biphasic constant current pulses that were applied between the selected finger electrode and the wrist electrode. A 30 dB range of EA was compressed into the range from threshold to "maximum comfortable level" for each electrode.

The sensations produced by the electrical stimulus, controlled by the speech processor, were such that loud sounds produced stronger sensations. Higher pitched voices produced higher pulse rates which were perceived as smoother sensations. The highest EF2 values caused stimulation by electrode 8 on the little finger while the lowest EF2 values caused stimulation by electrode 1 on the index finger. The subject's task was to interpret these dynamic patterns of stimulation as phonemes and words. Because this is not a naturally acquired skill, the proficiency of the subjects can be expected to be determined partly by their experience and previous training.

RESULTS

The mean scores obtained for the different tests in each condition are shown in Figure 1. For each test, a two-factor analysis of variance was carried out, using condition and subject as the factors. For the vowel, consonant and MTSP tests, the scores for separate blocks of data were used as repeated measures in the analysis. Similarly, word per minute scores for separate ten-minute sessions were used as repeated measures in the analysis of the tracking data. Since each subject was tested only once in each modality with the CNC words and BKB sentences, the mean square term for the interaction of subject and modality was used as the error term for the analysis. Every ANOVA showed a highly significant variation among the mean scores for the different modalities and their combinations.

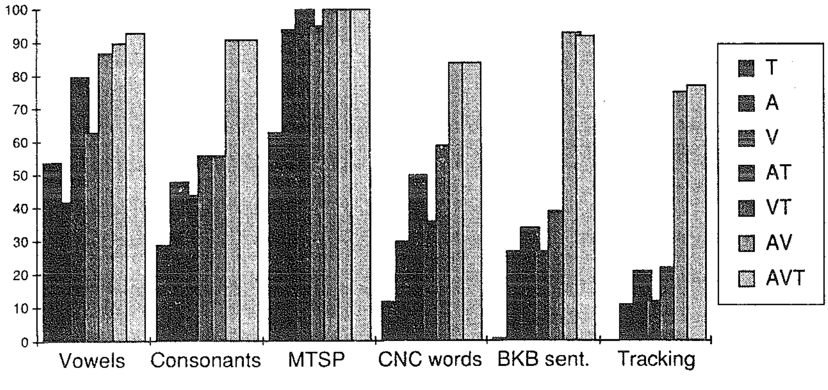

Figure 1. Mean percentage scores for each speech test in the different presentation conditions (N=4). Tracking scores are in words per minute instead of percent.

For each test, the mean scores for individual modalities and their combinations were compared using the Newman-Keuls procedure with a 95% confidence level for the criterion level. This led to the following relations between the scores:

For vowels:	A < T < AT < (V, VT) V < (AV, AVT)	[1]
For consonants:	T < (V, A) < (AT, VT) < (AV, AVT)	[2]
For MTSP:	T < (A, AT) < V	[3]
For CNC words:	T < (A, AT) < V < VT < (AV, AVT)	[4]
For BKB sentences:	T < (A, AT, V, VT) < (AV, AVT)	[5]
For tracking:	(A, AT) < (V, VT) < (AV, AVT)	[6]

In order to take a more detailed look at the combinations of cues provided by the three modalities, the vowel and consonant results were analysed in terms of the percentage of information transmitted for a number of features. This method was described in detail by Miller and Nicely (1955) who applied it to a study of auditory consonant confusions. The vowels were classified by duration, F1 and F2 on the basis of data for average male speakers given by Bernard (1970). Table 1 shows the percentage of information transmitted for each vowel feature in each condition, calculated from the confusion matrix for the four subjects together. In brackets after each percentage for a combined modality is the value predicted from the values for individual modalities. The prediction was made assuming that each modality independently contributed a proportion of the information and that an error occurred in the combined modality only if the speech feature was incorrectly perceived in both of the individual modalities. For example,

$$1 - I_{AV} = (1 - I_A)(1 - I_V) \qquad [7]$$

was used to predict the proportion of information transmitted in the AV condition, I_{AV}, from the proportion transmitted in the A and V conditions, I_A and I_V. In this case, the proportion of information incorrectly perceived is $(1-I_{AV})$ in the combined modality. In the individual modalities, the proportions of information incorrectly perceived are $(1-I_A)$ and $(1-I_V)$. Since the information provided by each modality is assumed to be statistically independent, the probability of an error in both modalities is the product $(1-I_A)(1-I_V)$. This formula has been shown to provide a good description of combined auditory-visual perception of nonsense syllables by cochlear implant users (Blamey *et al*, 1987). Similar equations were used to calculate predicted values for the AT, VT, and AVT combinations. Note that the values in Table 1 observed for the AV and AVT conditions were all greater than or equal to the predicted values. For the AT and VT conditions, the observed values were all less than the predicted values, with the exception of the proportion of duration information in the AT condition.

Condition			Feature	
	Total	Duration	F1 frequency	F2 frequency
A	43	91	33	24
V	75	71	76	75
T	43	49	28	51
AV	86 (86)	100 (97)	85 (84)	82 (81)
AT	56 (68)	96 (95)	40 (52)	50 (63)
VT	82 (86)	81 (85)	81 (83)	80 (88)
AVT	91 (90)	100 (99)	89 (88)	90 (89)

Table 1. Percentage of information transmitted for the vowels in the auditory, visual, tactile, and combined conditions. The values in brackets for the combined modalities are predicted from the values for individual modalities using equation [7].

Table 2 shows the percentage of information transmitted for each consonant feature in each modality, together with values for the combined modalities predicted from equation [7]. The first five features were used by Miller and Nicely (1955) and are based mainly on articulation of the consonants. The visibility feature is based on the three groups of consonants commonly distinguished by lipreaders. The last two features were used by Blamey *et al* (1987) to describe the information available to cochlear implant users in a similar experiment. They are based on the amplitude and F2 frequency parameters estimated by the speech processor. Observed scores for AV and AVT were greater than predicted scores with the exception of place for AVT. With the exception of affrication, observed VT scores were less than predicted. Observed AT scores were also less than predicted except for the nasality feature.

Feature				Condition			
	A	V	T	AV	AT	VT	AVT
Total	54	53	29	91 (78)	59 (67)	61 (67)	91 (85)
Voicing	79	7	11	91 (80)	78 (81)	15 (17)	94 (83)
Nasality	91	27	43	100 (93)	100 (95)	48 (58)	100 (96)
Affrication	66	70	29	95 (90)	70 (76)	87 (79)	96 (93)
Duration	38	58	80	94 (74)	74 (88)	81 (92)	95 (95)
Place	15	80	20	84 (83)	23 (32)	80 (84)	84 (86)
Visibility	24	100	12	100 (100)	29 (33)	100 (100)	100 (100)
Amplitude envelope	84	24	25	96 (88)	84 (88)	36 (43)	96 (91)
High F2	27	74	59	84 (81)	52 (70)	83 (89)	86 (92)

Table 2. Percentage of information transmitted for the consonants in the auditory, visual, tactile, and combined conditions. Values in brackets for combined modalities are predicted from the values for individual modalities using equation [7].

DISCUSSION

Firstly, it should be noted that each of the individual modalities A, V and T produced scores that were well above chance for the closed-set vowel, consonant and MTSP tests. In the case of the CNC words and BKB sentence tests with open response sets, non-zero scores were obtained for each modality. The scores for the T modality might be attributable to chance for these two tests. For vowel recognition, the T modality produced a higher score than the A modality, but for all other tests, T was the lowest scoring modality. V was the highest scoring unimodal condition for all tests except consonant recognition. The unimodal scores show that each modality was capable of conveying useful speech information, although the subjects were unable to use the tactile information effectively for open-set tasks.

Equations [1] to [6] show that the visual input was an effective supplement in every situation since VT > T, AV > A, and AVT > AT for every test. The auditory input also produced a significant increase, since AV > V, AT > T, and AVT > VT for every test except the vowels. In the case of the vowels, the third inequality did not reach the criterion at the 95% confidence level. This may have been a consequence of the limiting effect caused by the high score for VT (87%). The tactile input produced only four significant increases in score: AT > A for vowels; AT > A and VT > V for consonants; VT > V for CNC words. Despite the fact that the tactile input conveyed useful information in isolation, it did not seem to have as great an effect as A or V information when combined. This was especially true in the open-set tests.

Comparison of the observed and predicted information transmission values in Tables 1 and 2 also suggests that the AT and VT combinations were less effective than the AV combination in the closed-set tasks. The very effective combination of A and V may be a consequence of long experience with these modalities during the normal development of speech and language, especially in situations where the auditory signal is degraded by background noise. Alternatively, there may be specialized neural mechanisms for the combination of A and V information that are used in the AV mode of speech perception. The less effective combination of T information with either A or V may therefore be a consequence of lack of experience and appropriate training or lack of appropriate neural structures and functions to carry out the necessary combination. The data presented here are not sufficient to distinguish between these situations.

CONCLUSIONS

If it is assumed that there are no physiological limitations preventing the use of tactile information as effectively as auditory and visual information, the present study leads to several conclusions of practical significance. To be useful, a tactile aid must be capable of providing information in the T condition, and this information must be combined effectively with information from other modalities. The present study showed that it is possible to obtain a good score in the T modality without obtaining the full benefit of the tactile information in combined modalities. This implies that training in the combined modalities will be necessary, as well as training in the T condition. Also, the study showed that tactile information could supplement limited auditory information in closed-set tasks. Provided that training can extend this performance to open-set tasks, tactile devices may be beneficial to hearing aid and cochlear implant users with limited auditory function.

ACKNOWLEDGMENTS

Mr R. Cowan, Mr J. Alcantara, Ms L. Whitford, Ms I. Havriluk, Dr A. Blunden, Dr P. Seligman, Mr R. Millard and Mrs A. Brock provided support and advice that was essential to the evaluations carried out. At different stages of this research, financial support has been obtained from: the Deafness Foundation of Victoria, the National Health and Medical Research Council of Australia, the George Hicks Foundation, the Department of Industry, Technology and Commerce of the Australian Commonwealth Government and the Australian Bionic Ear Institute.

REFERENCES

Bench, J. & Bamford, J. (eds) (1979) *Speech-hearing tests and the spoken language of hearing-impaired children*, (Academic Press: London).

Bernard, J.R.L. (1970) *Toward the acoustic specification of Australian English*, Zeit. Phonetik 23, 113-128.

Blamey, P.J. & Clark, G.M. (1987) *Psychophysical studies relevant to the design of a digital electrotactile speech processor*, J. Acoust. Soc. Am. 82, 116-125.

Blamey, P.J., Dowell, R.C., Brown, A.M., Clark, G.M. & Seligman, P.M. (1987) *Vowel and consonant recognition of cochlear implant patients using formant-estimating speech processors*, J. Acoust. Soc. Am. 82, 48-57.

Cowan, R.S.C., Alcantara, J.I., Blamey, P.J. & Clark, G.M. (1988) *Preliminary evaluation of a multichannel electrotactile speech processor*, J. Acoust. Soc. Am. 83, 2328-2338.

de Filippo, C.L. & Scott, B.L. (1978) *A method for training and evaluating the reception of ongoing speech*, J. Acoust. Soc. Am. 63, 1186-1192.

Eilers, R.E., Widen, J.E. & Oller, D.K. (1988) *Assessment techniques to evaluate tactual aids for hearing-impaired subjects*, J. Rehab. Res. & Dev. 25, 33-46.

Erber, N.P. (1982) *Auditory training*, (A.G. Bell Association: Washington).

Miller, G.A. & Nicely, P.E. (1955) *An analysis of perceptual confusions among some English consonants*, J. Acoust. Soc. Am. 27, 338-352.

Peterson, G.E. & Lehiste, I. (1962) *Revised CNC lists for auditory tests*, J. Speech Hear. Dis. 27, 62-70.