Journal of Rehabilitation Research
and Development, Vol. 23, No. 1, 1986
pages 139-146

Lipreading with Tactile Supplements[a]

ARTHUR BOOTHROYD, Ph. D.
THERESA HNATH, M.A.

Center for Research in
Speech and Hearing Sciences
City University of New York
33 West 42nd Street
New York, N.Y. 10036

Abstract—Lipreading performance, after brief training, was measured in two normal subjects, as the percentage of words recognized in sets of unrelated sentences. Three receptive conditions were used: (i) lipreading alone, (ii) lipreading plus a single-channel tactile display of fundamental frequency (temporal only), and (iii) lipreading plus a multichannel, tactile, spatial display of fundamental frequency (temporal-spatial). After training, performance with tactile supplements was better than without, but it was not possible to conclude that either of the tactile displays was better than the other. During training, a significant correlation between performance and session number was found for the temporal-spatial display only.

INTRODUCTION

There is a long history of attempts to use the sense of touch as a substitute for hearing in the perception of speech. Most of that work has been aimed at alleviating the effects of profound or total sensorineural hearing loss by providing access to the information contained in the sound patterns of speech: e.g., Kirman, 1973 (1) and Levitt, Pickett, and Houde, 1980 (2).

In spite of early hopes to the contrary, the simple tactile transmission of the acoustic waveform via a single-channel transducer does not permit speech comprehension. It does, however, provide access to the amplitude envelope of the speech signal and can thus serve as a small but significant supplement to lipreading: e.g., Erber, 1972 (3); Gault, 1928 (4); Plant, 1982 (5); Zeiser and Erber, 1977 (6).

Tactile vocoders offer frequency resolution through a spatial transform of instantaneous spectral envelope. Their output has been shown to provide access to phonetic features and even to permit recognition of speech sounds and words in isolation (7, 8, 9, 10, 11, 12). However, performance on tasks requiring the perception of connected speech has been disappointing. At best, the multichannel tactile vocoder serves as a supplement to lipreading, providing an amount of help similar to that offered by a single-channel device: e.g., Sparks, Ardell, Bourgeois, Wiedmer, and Kuhl, 1979 (13).

Risberg, (14) pointed out that as long as tactile sensory aids are limited to the role of lipreading supplement they should be designed for the transmission of cues that are (i) inaccessible by lipreading and (ii) of major importance to speech perception. Intonation patterns are inaccessible by lipreading, and are important because they serve to

[a]The work described was supported in part by National Institutes of Health, NINCDS (National Institute of Neurological and Communicative Disorders and Stroke) Grant No. NS-17764.

define syntactic boundaries, to mark stress, and sometimes to distinguish statements from questions. In addition, access to fundamental voice frequency automatically provides information on voicing and on the durations and relative durations of voiced sounds. Experiments with normally hearing subjects have confirmed that the auditory presentation of intonation provides a dramatic supplement to lipreading: e.g., Risberg and Lubker, 1978 (15); Rosen, Fourcin, and Moore, 1981 (16). Such experiments suggest that a tactile intonation display might provide more lipreading assistance than has been demonstrated with single-channel displays of acoustic waveform.

Data on the frequency resolution of the skin suggest that a single-channel display in which fundamental frequency is encoded as the frequency of a pulse train should provide significant access to intonation contours: e.g., Rothenberg et al., 1977 (17); Rothenberg and Molitor, 1979 (18). Such a display is static, however, and does not take advantage of the specialized ability of the sense of touch to detect and process patterns of movement (1). In contrast, a spatial display, in which fundamental frequency is encoded as locus of stimulation, has the inherent advantage that intonation patterns become movement patterns. Such a display has been developed by one of us in connection with a study of sensory aids and their role in speech training with prelingually-hearing-impaired children (19). Similar devices have been described in the literature: e.g., Willemain and Lee, 1972 (20); Stratton, 1974 (21); and Grant, 1980 (22).

The present device, however, is wearable and therefore has the potential to function prosthetically to provide continuous input and feedback. Psychophysical studies have confirmed that the display provides a resolution of one stimulation channel, and that subjects can discriminate among contrasting intonation contours; see (23). The present study concerns the potential role of this device as a supplement to lipreading in the perception of connected speech.

The purpose of the study was to measure the recognition of words in sentences as a function of brief training, and to compare the scores obtained—

1. By lipreading alone;
2. By lipreading in combination with single-channel tactile presentation of fundamental frequency; and
3. By lipreading in combination with tactile presentation of the spatial transform of fundamental frequency.

The hypotheses were—

1. That recognition scores would improve with training;
2. That, after training, performance with tactile supplements would be superior to that with lipreading alone; and
3. That, after training, performance with the spatial display would be better than that found with the single-channel display.

EXPERIMENTAL PROCEDURE

Subjects

Two female graduate students in the Doctoral Program in Speech and Hearing Sciences at the City University of New York served as subjects. Both had normal hearing, and neither had previously had formal training in lipreading.

Instrumentation

The multichannel display employs an input transducer, a pitch extractor, a pitch-period meter, and, as output transducer, an eight-channel solenoid array (Fig. 1). Only one solenoid vibrates at a time, the locus being determined by fundamental voice frequency. As adjusted for this study, the frequency range covered by the eight channels was 80 to 350 Hz, channel number being proportional to the logarithm of frequency. The vibrating solenoid vibrates at half of the fundamental frequency of the input. Thus the encoding is both temporal and spatial. A more detailed description and a simplified schematic diagram of the electronic and electromechanical components will be found in the Appendix. Additional details appear in Reference 19.

This display will be referred to as the "temporal-spatial" display. For the purposes of the present study, the solenoid array was worn on the medial surface of the left (non-dominant) hand.

For the single-channel display, the device described above was modified so that there was the option of having only one solenoid activated, regardless of fundamental voice frequency. The frequency of vibration of that solenoid, however, was still half of the fundamental frequency of the input. For the rest of this report, the device thus modified will be referred to as the "temporal-only" display.

In order to prevent the subjects from hearing either the talker during training, or the acoustic output of the solenoids during training and testing, white noise was delivered to their ears, and the hand wearing the solenoid array was placed in a foam rubber tube (Fig. 2).

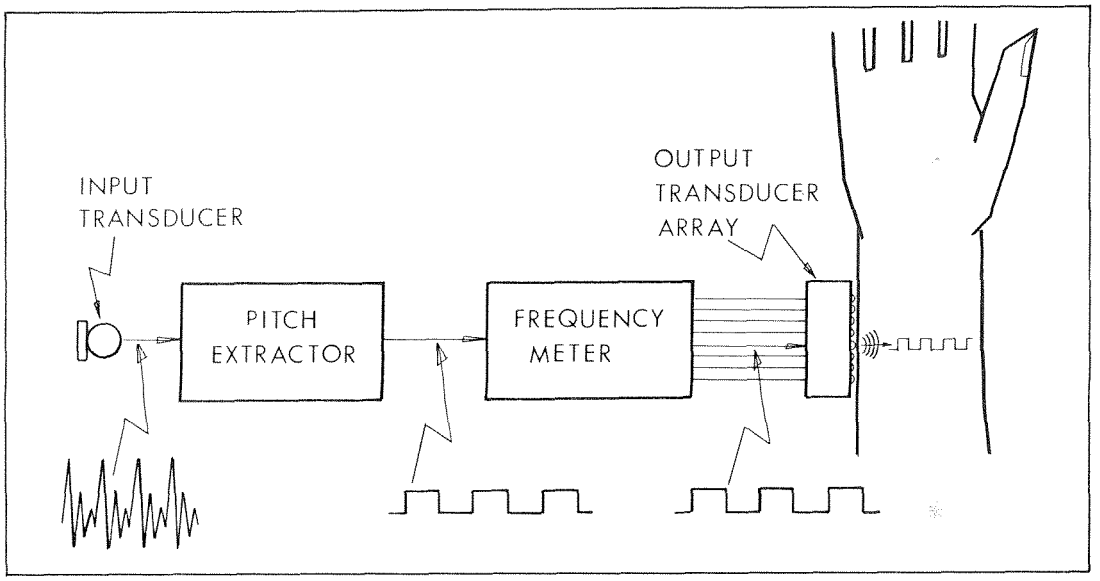

FIGURE 1
Block diagram of the temporal-spatial pitch display. Speech signals from the input transducer (an acoustic microphone or a surface-mounted accelerometer) pass to a pitch extractor that generates a square wave whose frequency equals one-half that of the fundamental voice frequency. A pitch period meter times the "off" period of each cycle and thereby determines which of eight output channels will be activated during the "on" period. Frequency scaling is such that the channel number is proportional to the logarithm of voice fundamental frequency.

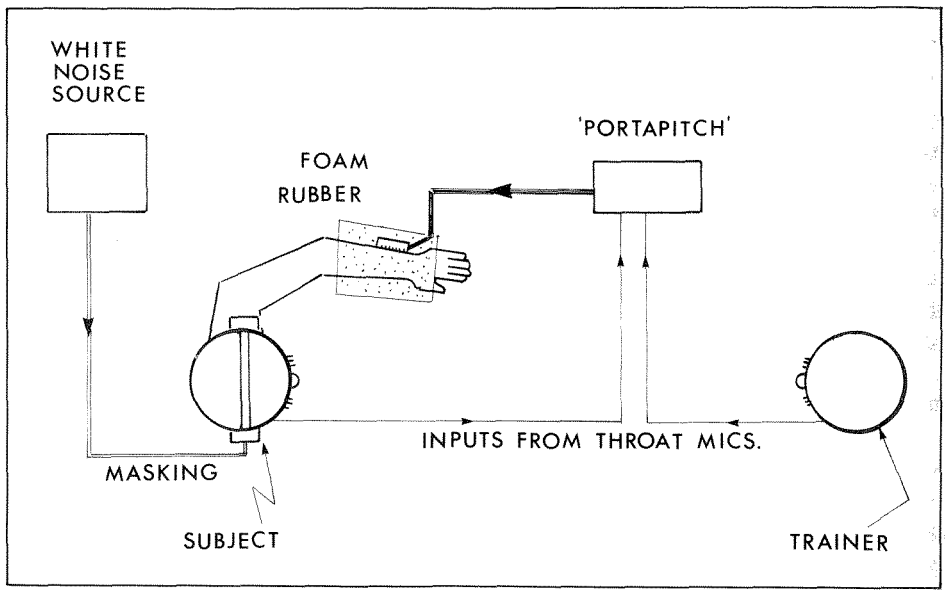

FIGURE 2
Plan diagram of the arrangement used for training of subjects. For formal testing, the trainer's input was replaced with the signal from one audio track of a videotape player. Note: In this diagram the solenoid array is shown being worn on the medial surface of the forearm, but the subjects in the present study actually wore it on the medial surface of the hand. "Portapitch" is the name the authors have given to the wearable, temporal-spatial display.

Formal tests of sentence perception were administered using video-recorded materials. One audio channel was recorded using an accelerometer attached to the talker's throat. During testing, the output from this channel was attenuated and connected to the microphone input of the tactile display.

Training Procedure

Training was carried out with live speech. Both the trainer and the trainee provided input to the tactile display from throat-mounted accelerometers. The trainer had a series of sentences written on index cards. Each sentence was read to the trainee who was required to repeat it. Her repetition was recorded on tape for later analysis. If the repetition was correct the next sentence was read. If the trainee could not repeat the sentence accurately, it was read a second time. If she still could not repeat the sentence, she was permitted to see it in printed form and a third opportunity for repetition was given before moving to the next sentence. This procedure was designed to satisfy several requirements:

1. The subject should move quickly through materials that are easy.
2. Long periods of frustration should be avoided.
3. The opportunity should be provided for learning to associate sensory input with motor output.

Test Materials

Test materials consisted of 200 sentences in 20 sets of 10 each. Within each set, sentence length varied uniformly from 3 to 12 words. Ten topics were represented (food, sports, clothing, etc.) and there was one sentence about each topic in each of the 20 sets. Each set contained at least 3 declaratives, 3 interrogatives, and 3 imperatives. The tenth sentence could be any one of the three types. Sentence length and type were randomized across sentence topics. These materials were video-recorded in color, with each sentence being spoken twice. Two audio tracks were recorded, one from an acoustic microphone located 6 inches from the mouth, the other from a throat-mounted accelerometer.

Method

Each subject was seen for six 1-hour sessions. The first session was used only for testing. In subsequent sessions, subjects received approximately 10 minutes of training (the time taken to complete 10 sentences) prior to each test under each of the three lipreading conditions (lipreading alone, lipreading plus the temporal-only display, and lipreading plus the temporal-spatial display).

A different set of 10 sentences was used for testing under each condition. The order of conditions was randomized across subjects and sessions. Different sentence sets were used for training and testing, and no sentence set was used more than once with a given subject.

During testing with the video-recorded material, subjects watched both productions of a sentence. Their spoken responses were then recorded on tape for subsequent analysis and scoring. Scores represented the number of words correct in a set of 10 sentences with each set containing 75 words.

It should be noted that the same talker was used for testing and training, there being no attempt to measure generalization of learned skills across talkers.

RESULTS

Table 1 shows the number of words recognized by each subject under each of the three lipreading conditions, during each of the six sessions. (These data are illustrated as percentage scores in Figure 3, which also includes linear regression functions for performance on session number.)

To test the hypothesis that performance would improve with training, linear correlations between performance and training-testing sessions were examined. There was a significant correlation for the composite score, summed across subjects and lipreading conditions: [$r(5) = 0.74$, p(single tail)$<.05$]. When the scores for the three lipreading conditions were examined separately, however, only the temporal-spatial condition showed significance: [$r(5) = 0.88$, p(single tail)$<.01$]. There was, therefore, strong evidence of a learning effect, most of which appeared to be attributable to the temporal-spatial condition.

The remaining hypotheses were tested using a three-way analysis of variance in the data from the last two sessions. The analysis, shown in Table 2, revealed a significant main effect for condition,

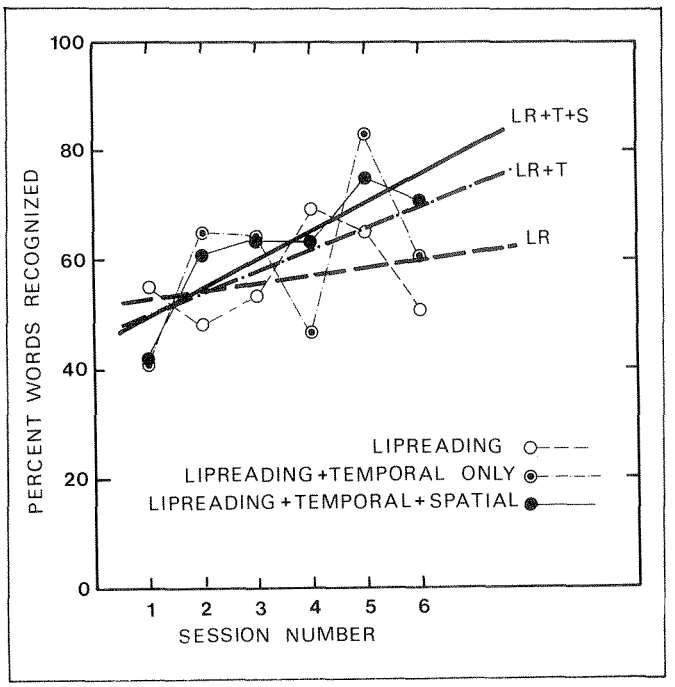

FIGURE 3
Average lipreading performance of two subjects as a function of training session. Data are shown for lipreading alone (LR), lipreading supplemented by the temporal-only display (LR + T), and lipreading plus the temporal-spatial display (LR + T + S). Heavy lines show linear regression functions for performance under each condition on session number. The coefficient of correlation for the temporal-spatial condition was significant at the 1 percent level [$r(5) = .88$, $p<.01$].

TABLE 1
Number of words recognized by each subject as a function of lipreading condition and training-testing session.

Condition	Subject	Session 1	2	3	4	5	6	Totals
Lipreading	A	35	33	41	57	64	44	274
	B	47	39	35	46	34	33	234
	A + B	82	72	76	103	98	77	508
Lipreading, + temporal	A	37	46	62	27	68	57	297
	B	25	52	34	44	57	34	246
	A + B	62	98	96	71	125	91	543
Lipreading, + temporal, + spatial	A	41	41	42	50	57	68	299
	B	22	51	52	45	56	39	265
	A + B	63	92	94	95	113	107	564
Totals	A	113	120	145	134	189	169	870
	B	94	142	121	135	147	106	745
	A + B	207	262	266	269	336	275	1615

[$F(2,2) = 20.01$, $p<.05$], supporting the hypothesis that performance was better with tactile supplements than without. Post hoc analysis, however, failed to reveal a significant difference between the two tactile conditions.

DISCUSSION

The data support the hypothesis that speechreading performance improves with brief training. They also support the hypothesis that (after brief training) speechreading with tactile supplements is better than without. The data do not, however, support the hypothesis that the temporal-spatial display provides more assistance than the temporal-only display.

Four major factors could have worked against the demonstration of significant effects in this study. One was the small amount of training in-

TABLE 2
Analysis of variance in the word-recognition scores.

Source of variance	Sum of squares	Degrees of freedom	Estimated mean square	Error term	F ratio
Subject	918.75	1	918.75		
Condition	310.75	2	155.09	SxC	20.01*
Replication	310.08	1	310.08	SxR	8.44
SxC	15.50	2	7.753	SxCxR	0.05
SxR	36.75	1	37.75	SxCxR	0.26
CxR	98.17	2	49.09	SxCxR	0.34
SxCxR	285.50	2	142.09		
Total	1974.92	11			

*$p<.05$

volved—a total of 1 hour per condition. A second was the mixing of conditions during each training session, a factor seen as likely to have caused interference among conditions. A third factor was the small number of subjects, leading to few degrees of freedom when examining the main effect of condition. The fourth was the high test-retest variability of the test scores. In an analysis of variance in the complete data set of Table 1, the residual sum of squares (SxCxR) is 1113.96, with 10 degrees of freedom, giving a standard deviation, for a single score, of 10.7 words. Thus, the 95 percent confidence limits for a single score are of the order of ±20 words. With so few subjects and so few training sessions, the demonstration of significance requires robust effects.

In spite of these constraints, the analysis of variance does provide evidence of benefit from the use of tactile supplements to lipreading, after a short period of training. If there are differences between the temporal-spatial and temporal-only supplements, longer training times, without mixing of conditions, would probably be required to demonstrate them.

The study did demonstrate evidence of learning in the form of a significant correlation between performance and session number. Of the three conditions, however, the temporal-spatial condition was the only one in which the effect reached statistical significance. In view of the small number of degrees of freedom and the correspondingly wide confidence limits of the correlation coefficients, the data do not justify the conclusion that the rate of learning was actually greater under the temporal-spatial condition than under the other two, although the results point in that direction and the analysis of variance supports the hypothesis.

It cannot be assumed that the formal test data were influenced by the training given between tests. Further studies would be required to determine the learning effects to be expected from repeated exposure to the test materials without feedback on performance.

An important condition omitted from the present study was that of lipreading supplemented by amplitude-envelope information from a single transducer. A complex tactile lipreading aid of the type tested in this study must not only demonstrate *significant improvements over lipreading alone,* but must prove itself superior to a simpler device. It is quite possible, for example, that the effects found in the present study were attributable solely to the access both displays provided to the presence and duration of voicing.

CONCLUSIONS

1. The results show significant evidence of short-term learning in the perception of connected speech when lipreading is supplemented by a temporal-spatial display of fundamental voice frequency.
2. The results do not, however, show significant evidence of learning when lipreading alone is used, or when a temporal-only supplement is used.
3. After training, subjects' lipreading performance was better with tactile supplements than without.
4. After training, performance with the temporal-spatial supplement was not significantly better than performance with the temporal-only supplement.
5. In spite of the small number of subjects and the limited amount of training, significant benefits were demonstrated from the use of tactile supplements to lipreading. Further research will be required to determine the limits of training and the possible benefits of the spatial representation of fundamental voice frequency■

APPENDIX

The temporal-tactile pitch display has four sections:

1. The input transducer,
2. The pitch extractor,
3. A pitch period meter, and
4. The output transducer array.

The input transducer can be an acoustic microphone or a surface accelerometer. The former is less intrusive, but the accelerometer provides a signal with higher energy at the fundamental frequency and is virtually immune to environmental noise.

The pitch extractor used in this system is very simple. A low-pass filter with a slope of between 12 and 18 dB-per-octave attenuates the second and higher harmonics of voiced sounds in relation to the fundamental. The resulting quasisnusoid is converted to a square wave, using a level-detector with hysteresis. A "flip flop" is then used to divide the frequency by 2 and to generate a square wave with a 50 percent duty cycle. The division by 2 brings frequencies within the optimal range of sensitivity for the skin and accommodates the response time of the output transducers. The 50 percent duty cycle is dictated by the design of the pitch period meter.

The pitch period meter uses a clock-driven shift register to measure the duration of the OFF portion of the input square wave, and by means of a set of AND gates, routes the ON portion to one of eight output channels. An exponentially rising voltage derived

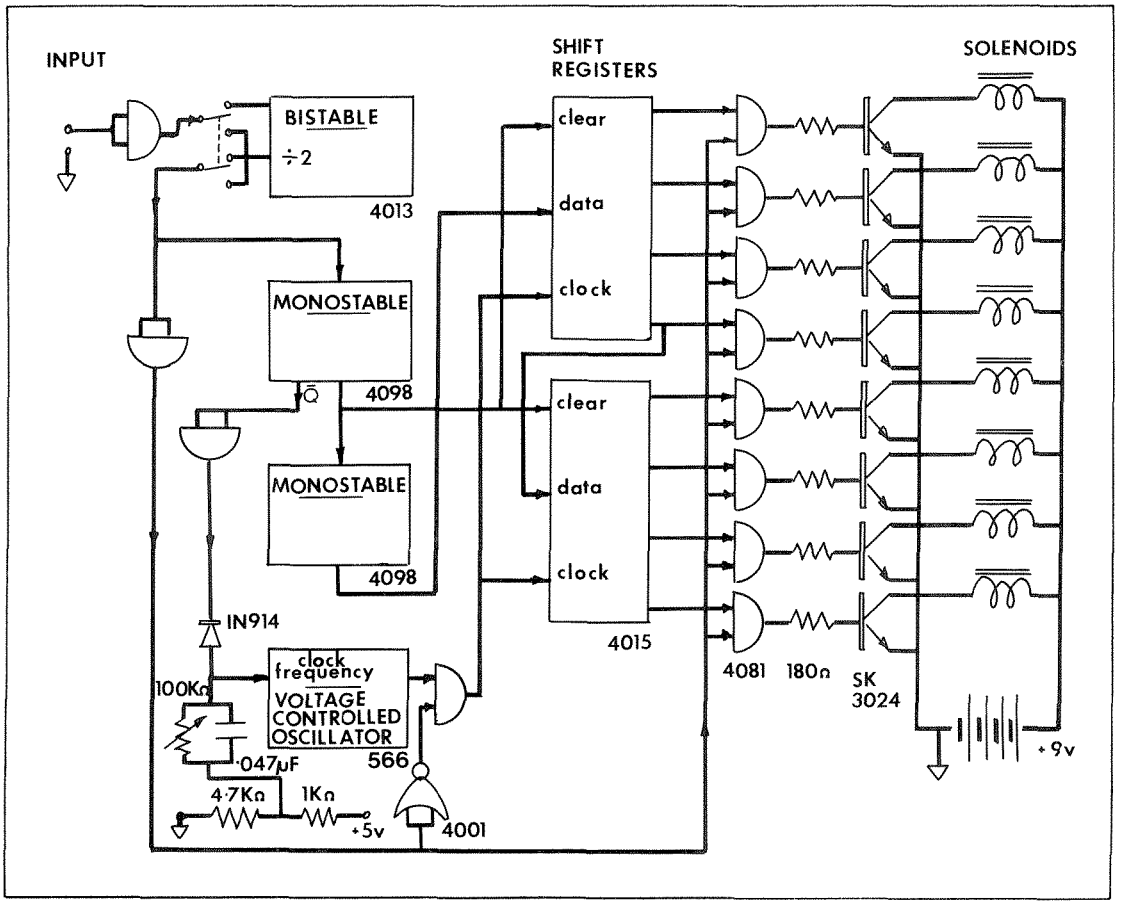

FIGURE A1.
Simplified schematic of the pitch period meter and output channels of the temporal-spatial pitch display.

from the CLEAR input to the shift register, modulates the clock so that channel number is proportional to the logarithm of the input frequency. A schematic of the pitch period meter is shown in Figure A1.

A set of transistors drives the output transducer array. This array consists of a set of eight miniature solenoids whose stroke has been reduced in length to approximately 0.5 mm. Current in the solenoid provides the closing force, and pressure from the skin provides the opening force. This arrangement is, in effect, a mechanical bistable system which provides a constant-amplitude square-wave stimulus to the skin when driven by a square-wave current.

REFERENCES

1. Kirman JH: Tactile communication of speech: a review and analysis. Psychol Bull 80:54-74, 1973.
2. Levitt H, Pickett JM, Houde RA: Sensory Aids for the Hearing Impaired. New York, IEEE Press, 1980.
3. Erber NP: Speech envelope cue as an acoustic aid to lipreading for the profoundly deaf. J Acoust Soc Am 51:1224-1227, 1972.
4. Gault RH: Interpretation of spoken language when the feel of speech supplements vision of the speaking face. Volta Rev 30:379-386, 1928.
5. Plant GL: Tactile perception by the profoundly deaf. Brit J Audiol 1982, 16:233-244, 1982.
6. Zeiser ML, Erber NP: Auditory/vibratory perception of syllabic structure in words by profoundly hearing-impaired children. J Speech & Hear Res 20:430-436, 1977.
7. Brooks PL, Frost BJ: Evaluation of a tactile vocoder for word recognition. J Acoust Soc Am 74:34-39, 1983.
8. Englemann S, Rosov R: Tactual hearing experiment with deaf and hearing subjects. J Exceptional Children 41:243-253, 1975.

9. Green BG, Craig JC, Wilson AM, Pisoni DB, Rhodes RP: Vibrotactile identification of vowels. J Acoust Soc Am 73:1766-1778, 1983.
10. Pickett JM: Tactile communication of speech sounds to the deaf: comparison with lipreading. J Speech Hear Res 28:315-330, 1963.
11. Pickett JM, Pickett BM: Communication of speech sounds by a tactile vocoder. J Speech Hear Res, 1963, 6:207-222, 1963.
12. Yeni-Komshian GH, Goldstein MH: Identification of speech sounds displayed on a vibrotactile vocoder. J Acoust Soc Am 62:194-198, 1977.
13. Sparks DW, Ardell LH, Bourgeois M, Wiedmer B, Kuhl PK: Investigating the MESA (Multipoint Electrotactile Speech Aid): the transmission of connected discourse. J Acoust Soc Am 65:810-815, 1979.
14. Risberg A: The importance of prosodic speech elements for the lipreader. **In:** H. Birk Nielsen and R. E. Klamp (Eds.), Visual and Audiovisual perception of Speech. Scandinavian Audiology, Supplement #4, 153-164, 1974.
15. Risberg A, Lubker JL: Prosody and speech reading. Stockholm: KTI Speech Transmission Laboratory, QPSR, No. 4, 1-16, 1978.
16. Rosen SM, Fourcin AJ, Moore BCJ: Voice pitch as an aid to lipreading. Nature 291:150-152, 1981.
17. Rothenberg M, Verillo RD, Zahorian SA, Brachman ML, Bolanowski SJ: Vibrotactile frequency for encoding a speech parameter. J Acoust Soc Am 64:1003-1012, 1977.
18. Rothenberg M, Molitor RD: Encoding voice fundamental frequency into vibrotactile frequency. J Acoust Soc Amer 66:1029-1038, 1979.
19. Boothroyd A: A wearable tactile intonation display for the deaf. IEEE Trans Acoust Speech & Sig Processing 33:111-117, 1985.
20. Willemain TR, Lee FF: Tactile pitch displays for the deaf. IEEE Transactions on Audio- and Electro-Acoustics AU-20, 9-16, 1972.
21. Stratton WD: Intonation feedback for the deaf through a tactile display. Volta Rev 76:26-35, 1974.
22. Grant KW: Investigating a tactile speechreading aid: the transmission of prosodic information in connected discourse and sentences. Masters Thesis, University of Washington, 1980.
23. Boothroyd A: Experiments with a wearable tactile pitch display. Proceedings of the International Conference on Rehabilitation Engineering, Ottawa, 1984.