Discrimination of Vowels with a Multi-finger Tactual Display

Ali Israr*†, Charlotte M. Reed‡ and Hong Z. Tan†

†Haptic Interface Research Laboratory, Purdue University, USA
‡Research Laboratory of Electronics, Massachusetts Institute of Technology, USA

ABSTRACT

The present study compared the performance of an ideal observer and a human participant in a vowel discrimination task using a speech-to-touch coding scheme designed for a three finger tactual display. The coding scheme extracted speech features and presented them as high-frequency vibrational and low-frequency motional waveforms. The high-frequency vibrations presented crude spectral information from three distinct speech bands on three fingerpads of the left hand. The same information was presented, redundantly, by the low-frequency waveforms. Performance of the ideal observer, where only high-frequency vibrational signals were considered, was evaluated by a signal detection theory using several tokens of a pair of vowels. Results showed that the acoustic cues corresponding to the first two formants were sufficient for discrimination of a seven vowel stimulus set. The participant was then tested in an absolute identification task with 640 tokens of ten non-diphthong vowels spoken by two female speakers. Both high- and low-frequency waveforms were presented to the participant. Discrimination scores for each pair of vowel were similar to the best scores obtained with the ideal observer indicating that the coding scheme was effective and the participant acted like an ideal observer.

CR Categories and Subject Descriptors: H.5.2 [Information Interfaces and Presentation]: User Interfaces – Haptic I/O; H.1.2 [Models and Principles]: User/Machine Systems – Human factors.

Additional Keywords: sensory substitution, vowel discrimination, tactual sense, detection thresholds, psychophysics.

1 INTRODUCTION

This work was motivated by our desire to use touch as a sensory substitute for hearing in speech communication. Previous investigators have developed artificial tactual communication systems using, for example, a simple single channel hand-held bone-vibrator [1, 2] or a multi-channel vibrating array of pins attached to the skin [3, 4] to transmit phonetic and prosodic features through the skin. Phoneme-level information transmitted through these systems, as evaluated by testing hearing and hearing-impaired participants, is available to some degree when such systems are used to supplement the visual cues provided by lipreading. In evaluations of systems where only tactile signals are available, however, overall speech performance is generally quite poor. In contrast, human ability to use the sense of touch for speech communication has been demonstrated in the natural (non-device based) method known as Tadoma, that has been used by deaf-blind individuals for speech communication [5]. In this method, the deaf-blind individual places his/her hand on the face of a speaker and monitors the mechanical signals that occur in speech production. Multidimensional articulatory cues such as lip opening, lip and jaw motion, laryngeal vibrations and air flow are sensed by the hands of the "listener" and can be used by experienced Tadoma users to understand the speech of both familiar and new speakers. Research has documented the remarkable capabilities of experienced Tadoma users to receive phoneme level as well as sentence level speech information. Tadoma users can receive oral speech at an information rate of 12 bits/sec – roughly half the rate at which normal conversations are conducted [6].

In comparison to the Tadoma method, the limited success of current artificial displays may be due in part to their limitation to the tactile (vibrational) sensory system (thus ignoring the kinesthetic component of the system that is present in Tadoma) and the effects of vibrotactile masking [7-9]. Current displays are useful in transmitting crude acoustical features, such as duration, modulation, periodicity, etc., but not fine spectral variations within speech segments. One possible approach for improving the transmission of speech through the skin is through the utilization of both the kinesthetic and tactile components of the sense of touch. It is well known that the low-frequency motional and high-frequency vibrational waveforms stimulate two independent sensory mechanisms that do not interfere with each other at threshold and suprathreshold levels [10-12].

The TACTUATOR, a multi-finger tactual stimulator, was developed to broaden the dynamic range of tactual stimulation by delivering multidimensional waveforms to the fingerpads of the hand [7]. Previous experiments using synthetic signals with the TACTUATOR have demonstrated an information rate of 12 bits/sec, which is roughly comparable to that achieved with the natural Tadoma method [13]. A second system (called TACTUATORII) was subsequently developed with a new two-degree-of-freedom controller that preserves the relative intensities of the spectral components in the input signals in terms of perceived intensities in sensation levels [14]. A speech-to-touch coding scheme has been developed for TACTUATORII that extracts acoustic features from recorded speech segments and presents them as high-frequency vibrations and low-frequency motional waveforms through the three channels of TACTUATORII. The high-frequency vibrations present the overall spectral features extracted from three frequency bands of the speech spectrum to the thumb, index and middle fingers of the left hand. The low-frequency motional waveforms map the finer variations of the corresponding frequency band to each finger channel, presenting the spectral information redundantly.

In this paper, the coding scheme was evaluated by measuring the performance of an ideal observer in a vowel-discrimination task using the signals presented through the vibrational display only. The performance of the ideal observer was then compared to experimental data obtained from a human participant in a vowel-identification task through the tactual display. Signal detection theory was used to calculate discrimination scores [15, 16]. Discrimination performance of the ideal observer was evaluated by identifying a set of viable acoustic cues and determining the probability distribution functions of each cue for

* email: israr@rice.edu. Currently working as a research associate in Mechatronics and Haptic Interface Laboratory at Rice University, USA.

Symposium on Haptic Interfaces for Virtual
Environments and Teleoperator Systems 2008
13-14 March, Reno, Nevada, USA
978-1-4244-1972-2/08/$25.00 ©2008 IEEE

multiple tokens of a pair of vowels spoken by a single speaker. Sensitivity indices for each pair of vowels were determined by assuming that the distributions were normal (Gaussian) with the same variances. Multiple tokens of ten “pure” (non-diphthong) vowels spoken by two speakers were identified by the participant in three experimental conditions (pre-training, training, and post-training) in order to determine the effects of training in the vowel identification task. The performance levels of the ideal observer and the human participant were then compared.

The remainder of the paper is organized as follows. In Section 2 the speech-to-touch coding scheme is presented. Section 3 presents the vowel discrimination analysis of the ideal observer using the proposed coding scheme. Section 4 presents methods and results of the vowel identification experiment with a human participant. The paper concludes with a general discussion in Section 5.

2 Speech-to-touch Coding Scheme

The speech-to-touch coding scheme was specifically-designed for the TactuatorII display. TactuatorII consists of three single-degree-of-freedom actuators that interface the middle finger, the index finger, and the thumb. The selection of signal-processing strategies and coding schemes for the tactual speech display was guided by previous research in a number of areas, including, for example, auditory speech perception, previous work on speech reception through tactual displays, methods of communication employed by deaf-blind individuals, and tactual psychophysics. Due to limited space, we can not elaborate on the justification of each parameter, but have tried to cite relevant studies in these areas which led to the specific choices employed in our system.

For extraction of spectral features from the recorded speech material, the following schemes were used in Matlab: 1) low-pass filter, 2) band-pass filter and 3) envelope extraction scheme. Figure 1 shows a block diagram of the amplitude envelope extraction scheme as presented in [16, 17]. In this scheme, a band-limited signal was rectified and passed through a low-pass smoothing filter (6th-order Butterworth) in order to extract the temporal envelope of the input signal. The envelope was then scaled and outputted with a carrier frequency F_c.

2.1 Vibrational Coding

Figure 2 shows the block diagram of the extraction scheme and illustrates the transformation of extracted speech information into vibrotactile waveforms (mid- and high-frequency waveforms). Spectral information from three distinct bands (F0-, F1- and F2-bands) was presented through the three channels of TactuatorII. The speech signal was first passed through a pre-emphasis filter that amplified the energy above 1000 Hz at the typical rate of 6 dB per octave. Fundamental frequency (F0) information was presented at the thumb channel by passing the low-pass filtered speech signal directly through the 2-dof controller of [14] (see route 1 in Figure 2). Spectral information of the first formant (F1) and second formant (F2) was presented through the middle finger channel (route 2 in Figure 2) and through the index finger channel (route 3 in Figure 2), respectively. Through route 2, the pre-emphasized signal was passed through a band-pass filter that had a pass band in the first formant (F1) frequency region. The band limited signal was further processed through two band-pass filters. The amplitude envelopes of these two bands were extracted and then modulated with carrier frequencies of 30 and 200 Hz, respectively. The 30 Hz carrier modulated the envelope of the lower band and the 200 Hz carrier modulated that of the higher band. The two vibrational signals were added and presented through the middle finger channel. A similar scheme was used to extract envelopes in the second formant (F2) region and presented to the index finger channel (route 3 in Figure 2).

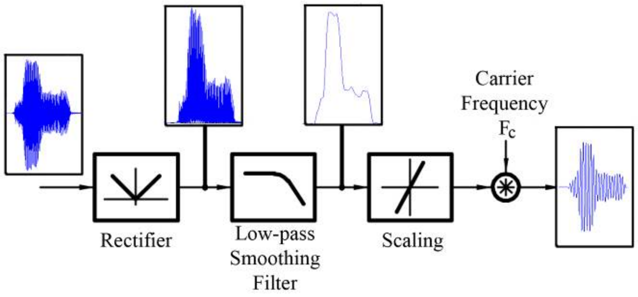

Figure 1. Envelope extractor

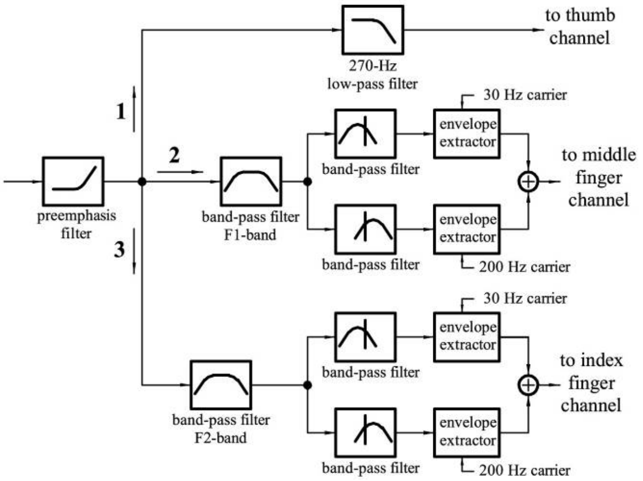

Figure 2. Block diagram of vibrational coding scheme

All envelopes were scaled by a straight line function with a slope of 0.5 (dB SL/dB SPL) and an intercept of 40 dB SL in order to cover the tactile dynamic range but not to cause pain or discomfort [18]. Since the digitized speech segments were normalized to one, the vibrations were scaled to a maximum intensity of 40 dB SL. The cut-off frequencies of the F0, F1 and F2 frequency band, and the lower and higher bands in the F1 and F2 frequency band are shown in Table 1.

2.2 Motional Coding

The motional coding scheme kept track of the frequency of the largest spectral peak in each finger band and encoded these spectral peaks as low-frequency (< 8 Hz) motion cues. These cues indicated variations of spectral energy as well as the frequency value of the spectral peaks in each finger band. Figure 3 shows a block diagram of the signal processing schemes used for extracting frequency variations and mapping them as low-frequency motional cues through all three channels. As before, routes 1, 2 and 3 corresponded to the features presented at the

Table 1. Speech bands and corresponding vibrations

Finger Channel	Speech bands (Hz)	Envelope bands (Hz)	Carrier frequency (Hz)
Middle finger	F1 band (300-1200)	300-650	30
		650-1200	200
Index finger	F2 band (1150-4000)	1150-1750	30
		1750-4000	200
Thumb	F0 band (80-270)	Low-pass filtered at 270 Hz	

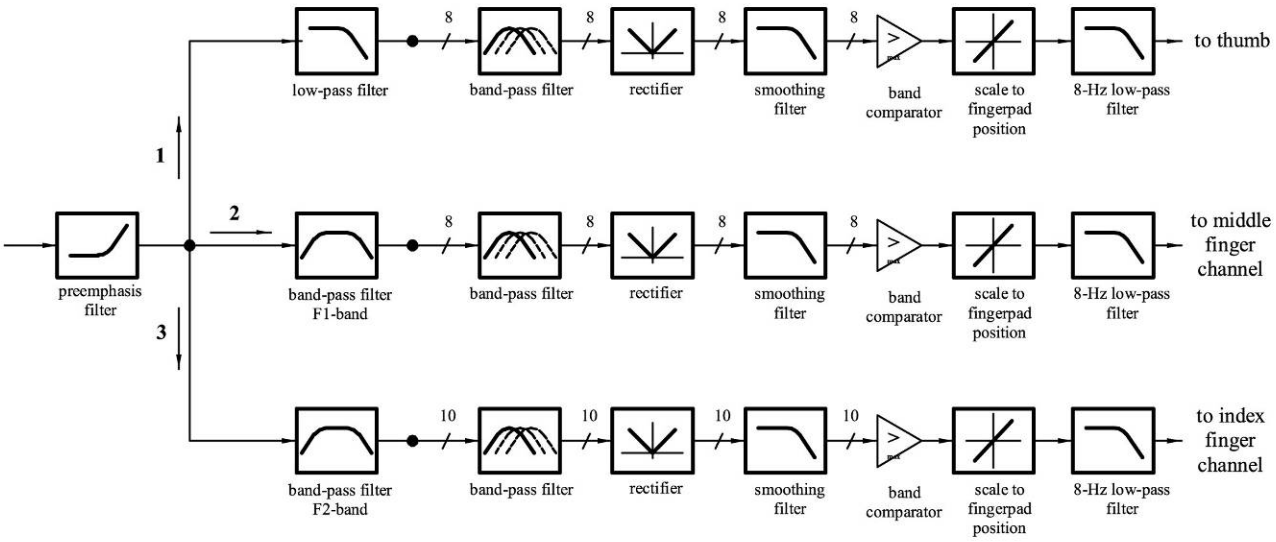

Figure 3. Block diagram of motional coding scheme

thumb, the middle finger and the index finger channels, respectively. Through route 1, the low-pass filtered signal was passed through eight contiguous band-pass filters in parallel and temporal envelopes of each band were evaluated. All eight envelopes were compared in a comparator and the center frequency of the band with the largest envelope value was noted at each sample instant. The center frequency was mapped to the absolute reference position of the fingerpad interface that ranged ±12.5 mm in its motion.

Similar schemes were implemented on route 2 and 3 corresponding to the frequency variations of the F1 and F2 peaks through the middle finger channel and the index finger channel, respectively. The center frequencies and bands of each band-pass filter are shown in Table 2. The frequency bands of the middle finger and thumb channels were divided into eight bands, while the frequency band of the index finger channel was divided into ten bands in order to cover its larger frequency range.

3 An Ideal Observer

The performance of an ideal observer in vowel discrimination was measured in a manner similar to that described in [16]. Treating the acoustic cues described above for vibrational coding as perceptual-distance in a decision space, sensitivity indices were calculated as measures for vowel distinction using signal-detection theory [15]. [Note that only vibrational cues were considered in this analysis.]

Table 2. Frequency bands for motional cues

Filter index	Middle finger	Index finger	Thumb
	Frequency band (Hz)		
1	300 – 400	1150 – 1300	80 – 100
2	400 – 500	1300 – 1500	100 – 120
3	500 – 600	1500 – 1700	120 – 140
4	600 – 700	1700 – 1900	140 – 160
5	700 – 800	1900 – 2100	170 – 200
6	800 – 900	2100 – 2300	200 – 220
7	900 – 1000	2300 – 2500	220 – 240
8	1000 – 1200	2500 – 3000	240 – 260
9	N/A	3000 – 4000	N/A
10	N/A	4000 – 5000	N/A

3.1 Speech Material

The speech material consisted of a subset of the Consonant-Vowel-Consonant (C1-V-C2) nonsense syllable database described in [16]. The speech material consisted of twenty tokens of seven vowels (/ae, ah, ee, eh, ih, oo, uu/) selected from the C1VC2 syllables spoken by one female speaker and stored as .mov files. The tokens corresponded to twenty syllables with C1 selected from /p, t, k, b, d, g, f, th, s, sh, v, tx, z, zh, ch, j, m, n, r, l/. Adobe Premiere 6.0 (Adobe Systems Inc., San Jose, CA) was used to separate the V segments from the C1VC2 segments. Care was taken to eliminate the transitional cues between V and C2, and between C1 and V. First, C2 was clipped from the C1VC2 nonsense syllables. Second, C1 was clipped from the resulting C1V segments to obtain the 140 vowel-only segments that were saved as .mov files. The .mov files were then converted into .wav (waveform audio) files using ConvertMovie 3.1 (MOVAVI, Novosibirsk, Russia) with the audio format set at a sampling rate of 11,025 Hz and 16-bit mono.

3.2 Acoustic Cues

Many studies have reported that vowels can be classified on the basis of the first two formants (see, for example, [19, 20]). Hillenbrand et al. [21] measured the acoustic characteristics of American English vowels spoken by 150 men, women and children and concluded that vowels could be discriminated with a high accuracy if duration and spectral change information was included in addition to formant patterns. Thus, the following four acoustic cues were measured from the listed seven vowels: 1) intensity of the fundamental frequency, A0; 2) duration of the vowel, D; 3) the relative amplitude of vibrations through the middle finger channel (F1 band), ΔA_1; and 4) the relative amplitude of vibrations through the index finger channel (F2 band), ΔA_2. A0 was calculated as the average of the spectrum in the band of 0-270 Hz. Relative amplitudes are defined as follows:

$$\Delta A_i = \text{mean}\left[(En)_{30Hz} - (En)_{200Hz} \right] \quad (1)$$

where (En) represents the envelope, i=1 for the middle finger channel and i=2 for the index finger channel.

3.3 Data Analysis

In a decision (or perceptual) space, shown in Figure 4, *x* is defined as the random variable along a decision axis representing stimulus cue and p(*x*|S) is the conditional probability density function of the cue given stimulus S.

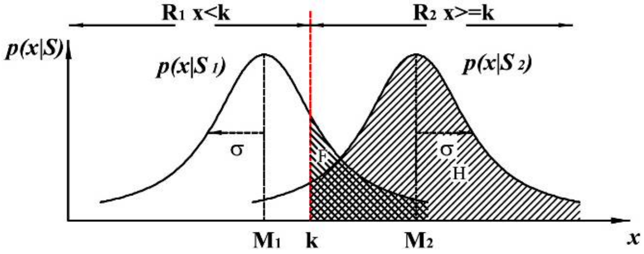

Figure 4. A decision space

The sensitivity index is defined as,

$$d' = \frac{M_2 - M_1}{\sigma} \tag{2}$$

where M_i is the mean of the Gaussian distribution for stimulus S_i; $\sigma = \sigma_1 = \sigma_2$ = standard deviation of the distributions of stimulus S_1 or S_2, R_i the correct response to S_i, and k the response criterion. Since d′=1 is usually used as the performance criterion for discrimination threshold, d′ > 1 indicates that two stimuli are discriminable.

The distributions obtained by a one-interval two-alternative (1I-2A) procedure were not used in determining the sensitivity index because preliminary inspection of the distributions did not meet the equal variance requirement. Instead, a two-interval two-alternative (2I-2A) procedure for pair-wise discrimination was employed because it yields equal variances for the probability density functions [15]. Sensitivity index is calculated by Eq.2 and converted to its equivalence for the 1I-2A procedure by the following formula

$$d'_{1I} = \frac{1}{\sqrt{2}} d'_{2I} \tag{3}$$

where d'_{1I} and d'_{2I} represent the sensitivity indices for the one-interval and the two-interval procedures, respectively [15].

3.4 Procedure

The procedure to determine distribution of a specific cue from a 2I-2A procedure for a pair of vowels is as follows:

1) Randomly select a segment from the 20 tokens of the first vowel.
2) Randomly select a segment from the 20 tokens of the second vowel.
3) Measure the acoustic cues from both segments and calculate the difference of the cues. Store the difference in an array.
4) Repeat steps 1 to 3 *n* times, where *n* >1000.
5) Divide the complete range of the array into 50 equal size bins and count the number of tokens in each bin.
6) Determine the proportion of occurrences within each bin by dividing the number of tokens in each bin by the total number of tokens *n*.
7) Determine the cumulative probability by adding the proportion of occurrences of a bin to those in the bins to the left.
8) Plot the cumulative probability densities of all 50 bins against the acoustic cue.
9) Now reverse the order of the pair and repeat steps 1 to 8.

3.5 Results

Sensitivity indices equivalent to a 1I-2A procedure of all combinations of seven vowel pairs are shown in Table 3 for A0, in Table 4 for D, in Table 5 for ΔA_1 and in Table 6 for ΔA_2. Sensitivity indices greater than 1 are highlighted in order to indicate the discriminable cues between each pair. The average (and standard deviation) of sensitivity index (d′) over the 21 pairs of vowels for A0, D, ΔA_1 and ΔA_2 are 1.63 (1.18), 0.93 (0.64), 4.19 (3.44) and 3.17 (1.94), respectively. The largest d′ occurred for acoustic cues corresponding to the first two formants, i.e. ΔA_1 and ΔA_2.

Table 3. d′ for acoustic cue A0

	/ae/	/ah/	/ee/	/eh/	/ih/	/oo/	/uu/
/ae/		0.3	**3.2**	0.5	0.9	**4.1**	**1.5**
/ah/			**3.1**	0.8	**1.1**	**3.8**	**1.6**
/ee/				**2.4**	**1.1**	0.6	**1.1**
/eh/					0.6	**3.1**	**1.0**
/ih/						**1.6**	0.3
/oo/							**1.5**
/uu/							

Table 4. d′ for acoustic cue D

	/ae/	/ah/	/ee/	/eh/	/ih/	/oo/	/uu/
/ae/		0.8	**1.1**	**1.1**	**2.2**	0.7	**2.6**
/ah/			0.4	**1.0**	**1.4**	0.1	**1.6**
/ee/				0.6	0.9	0.2	**1.0**
/eh/					0.3	0.8	0.5
/ih/						**1.0**	0.1
/oo/							**1.2**
/uu/							

Table 5. d′ for acoustic cue ΔA_1

	/ae/	/ah/	/ee/	/eh/	/ih/	/oo/	/uu/
/ae/		0.7	**6.6**	0.8	**4.6**	**7.0**	**4.5**
/ah/			**11.6**	**1.6**	**7.0**	**12.3**	**7.0**
/ee/				**4.8**	**1.2**	0.6	**1.8**
/eh/					**3.4**	**5.2**	**3.2**
/ih/						**1.5**	0.4
/oo/							**2.2**
/uu/							

Table 6. d′ for acoustic cue ΔA_2

	/ae/	/ah/	/ee/	/eh/	/ih/	/oo/	/uu/
/ae/		**3.1**	**7.0**	**2.8**	**3.1**	**1.2**	**1.5**
/ah/			**7.2**	**4.6**	**4.8**	**1.6**	0.9
/ee/				**3.1**	**1.3**	**5.6**	**5.0**
/eh/					**1.0**	**2.9**	**2.9**
/ih/						**3.3**	**3.3**
/oo/							0.4
/uu/							

4 VOWEL IDENTIFICATION EXPERIMENT

4.1 Apparatus

Multidimensional tactual waveforms were presented through the TACTUATORII. The three channels of the display corresponded to the three point contacts with the middle finger, the index finger and the thumb. Each channel was capable of delivering low-frequency motional cues and high-frequency smooth vibrations as well as mid-frequency flutter waveforms. More detail of the display is presented in [12, 14]. The participant's left hand was used to receive the tactual signals.

4.2 Speech Material

The speech material consisted of a subset of nonsense syllables spoken by two female speakers and described in [16]. The syllables were converted into digital segments and stored as .mov files. Each file started 2-4 frames before the initial lip opening and stopped 2-4 frames after the final lip closure. The syllables consisted of 4 tokens (2 tokens per speaker) of 16 C1 (/p, t, k, d, g, m, f, th, v, tx, ch, j, r, w, l, h/) joined by 10 non-diphthong (or "pure") medial vowels (/ae, ah, aw, ee, eh, er, ih, oo, uh, uu/). The C2 was randomly selected from a set of 21 consonants in the C1VC2 format segments. The duration of the segments varied from 0.868 to 2.502 sec with a mean of 1.574 sec. One half of the segments (16 C1 × 10 V × 1 tokens × 2 speakers = 320 segments) were used in the training sessions and the other half were used in the test sessions. The .mov files were then converted into .wav files with audio settings similar to those described in Sec.3.

4.3 Procedure

One male participant (S1, 30 years old, who is also a co-author of this study) took part in the experiment. The participant was highly experienced with the TACTUATORII device and the multidimensional cues presented through the device. He sat in front of the computer screen and followed instructions displayed on the screen to start the experiments. The vowel-identification experiment was conducted using a one-interval ten-alternative forced-choice (1I-10AFC) paradigm. On each trial, the participant was presented with a stimulus (vibrational and motional waveforms) corresponding to a randomly selected vowel from the token set. He was instructed to respond by pressing a button corresponding to the vowel presented using the mouse with his right hand. After the response, a new trial began. The duration of each stimulus interval was set to 2 seconds. A Hanning window (50-msec rise and fall time) was incorporated in order to eliminate the abrupt onset and offset of the stimulus. Each experimental run involved 50 trials that lasted for about 6-9 minutes. All stimuli started and ended at the mid-point of the range of motion of the three channels.

The participant was tested in three experimental conditions, pre-training test, training and post-training test conditions. The pre-training test condition consisted of two experimental runs of 50-trials each. The training condition consisted of 50 50-trial runs where trial-by-trial feedback was provided to the participant. The correct answer feedback was displayed on the computer screen at the end of the trail, after the participant's response. The post-training test condition consisted of 10 50-trial runs. No correct answer feedback was provided in the pre- and post-training test conditions. At the end of each testing and training run, the percent-correct score of the run was displayed on the computer screen. The participant was tested for no more than two hours in a day.

During the experiments, the TACTUATORII was placed to the left of the participant's torso. It was covered by a padded wooden box that served as an armrest for the participant's left forearm. The top of the box had an opening through which the participant could reach in and rest the thumb, index and middle fingers on the corresponding actuators. Pink noise (presented through circumaural headphones at roughly 80 dB SPL) was used to eliminate possible auditory cues. The participant was given a sheet of paper showing the spectral features (F1-F2 space and intensities of each vowel) associated with all 10 vowels. The participant was encouraged to use the paper in the pre-training test and training conditions.

4.4 Data Analysis

The results of the vowel identification experiment were expressed in terms of information transfer (IT) as in [22]. A 10×10 stimulus-response confusion matrix was formed for each run and each experimental condition. The trials with the same vowel in a C1VC2 pair were pooled together, so that the 10 stimulus alternatives corresponded to the ten vowels: S_1 = /ae/, S_2 = /ah/….. S_{10} = /uu/. Accordingly, ten responses were: R_1 = /ae/, R_2 = /ah/….. R_{10} = /uu/. The maximum likelihood estimate of IT was calculated by using

$$\mathrm{IT_{est}} = \sum_{j=1}^{k}\sum_{i=1}^{k}\frac{n_{ij}}{n_i n_j}\log_2\left(\frac{n_{ij}n}{n_i n_j}\right) \qquad (4)$$

where k = 10 was the number of stimulus alternatives, n was the total number of trials, n_{ij} was the number of times the joint event (S_i,R_j) occurred, and $n_i = \sum_{j=1}^{k} n_{ij}$ and $n_j = \sum_{i=1}^{k} n_{ij}$ were the sum of trials for each row and column, respectively. The percentage-correct scores (PC) were calculated by using

$$\mathrm{PC} = \sum_{i=1}^{k}\frac{n_{ii}}{n} \qquad (5)$$

The IT and PC were compared in pre- and post-training conditions to highlight training effects. In order to compare the results of the vowel-identification experiment with the performance of the ideal observer, 2 × 2 stimulus-response confusion matrices were formed for each vowel pair. Sensitivity index for each pair was calculated by using

$$d' = z(H) - z(F) \qquad (6)$$

where the hit rate, H = N(hits)/[N(hits)+N(misses)], is the proportion of responding R_2 when S_2 was presented. The false-alarm rate, F = N(false alarms)/[N(false alarms)+N(correct rejections)], is the proportion of responding R_2 when S_1 was presented. z(.) is the inverse of a normal (Gaussian) distribution function [15]. The sensitivity index was saturated at 4.65 corresponding to the percentage correct score of 99%.

4.5 Results

The percentage correct (PC) scores in training runs and testing runs are shown in Figure 5.

In training runs, the PC scores increased with experimental runs and the performance did not reach saturation after 50 runs. The PC scores improved from 16% in run 2 to 66% in run 43. A straight line was regressed through the PC scores along the experimental runs resulted in a significant slope ($p<0.001$) and $r^2=0.77$. Overall, the performance (both IT and PC) increased in the post-training condition as compared to the pre-training test

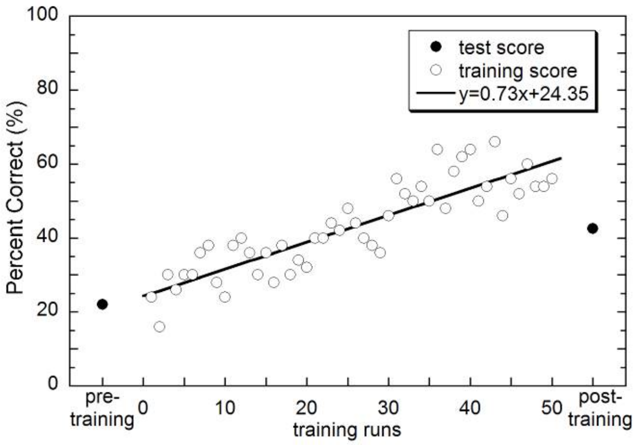

Figure 5. Percentage correct scores in identification experiment

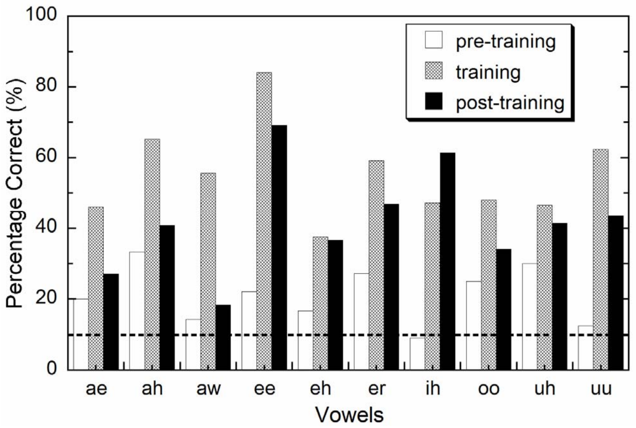

Figure 6. Percentage correct scores for vowels

condition. The PC scores increased from 22% in pre-training testing to 42.6% in post-training, where the chance level was 10%. The IT increased from 0.86 bits in pre-training test to 1.16 bit in post-training test condition. The performance level in the last 20 runs of training was higher than that obtained in the post-training condition both in terms of PC and IT (PC = 55.3% and IT = 1.32 bits). This result suggests both that (a) the trial-by-trial correct-answer feedback was useful to the participant in performing the identification task and that (b) some of the learning accomplished in the training sessions may have been specific to the speech tokens used in training, to which the subject had repeated exposure over the course of the 50 runs of training.

The ability to correctly identify each vowel in pre-training, during the last 20 runs of training and in post-training conditions, is presented in Figure 6. The dashed horizontal line indicates the chance performance level of 10%. In pre-training, PC varied from 9.1% for /ih/ (roughly chance level) to 33.3% for /ah/. For the runs in training, PC varied from 37.5% for /eh/ to 84.1% for /ee/. In post-training, the PC values were well above the chance level for all the vowels and varied from 18.4% for /aw/ to 69.2% for /ee/. For all the vowels, the performance scores were least in pre-training and largest in training except for the vowel /ih/ that was identified best in post-training test condition.

In order to compare the results of the identification experiment with the performance of the ideal observer, sensitivity indices for training and post-training test scores were calculated for each pair of vowels. The sensitivity indices for ten pure vowels are presented in Table 7 for last 20 runs of training and in Table 8 for post-training conditions. In general, sensitivity indices in the last 20 runs of training were comparable to those in post-training testing. The average d′ over the 45 pair of vowel was 3.26 (std 1.1) in training and 2.94 (std 1.35) in post-training testing. A two-sided t-test for the two condition showed non-significant effects (p>0.05) in performance in the two conditions.

Table 7. d′ for vowel pairs in the last 20 runs of training

	ae	ah	aw	ee	eh	er	ih	oo	uh	uu
ae		**1.5**	**1.5**	**4.7**	**3.6**	**3.7**	**4.1**	**4.4**	**2.0**	**4.5**
ah			**1.7**	**4.7**	**3.4**	**4.1**	**4.2**	**4.7**	**3.2**	**4.5**
aw				**4.7**	**3.8**	**3.8**	**4.7**	**4.7**	**2.5**	**3.7**
ee					**2.5**	**4.5**	**2.4**	**2.8**	**4.5**	**4.7**
eh						**1.6**	**1.3**	**2.6**	**1.3**	**2.7**
er							**2.9**	**2.9**	**2.4**	**3.2**
ih								**2.1**	**3.4**	**2.4**
oo									**3.0**	**2.0**
uh										**3.2**
uu										

Table 8. d′ for vowel pairs in post-training

	ae	ah	aw	ee	eh	er	ih	oo	uh	uu
ae		0.1	0.2	**4.7**	**2.6**	**3.2**	**4.7**	**3.8**	**1.3**	**4.7**
ah			-0.2	**4.7**	**3.9**	**3.4**	**4.7**	**3.8**	**2.6**	**4.1**
aw				**4.7**	**2.3**	**3.6**	**4.7**	**3.6**	**1.1**	**2.0**
ee					**3.9**	**3.8**	**2.3**	**2.6**	**4.7**	**3.6**
eh						**1.8**	**2.5**	**2.3**	**1.4**	**4.0**
er							**3.1**	**2.6**	**2.0**	**1.0**
ih								**3.4**	**2.8**	**2.4**
oo									**3.8**	**1.2**
uh										**3.0**
uu										

5 Discussion

In this paper, vowel discrimination and identification performance was evaluated for an ideal observer and a human participant using a speech-to-touch coding scheme proposed for the TACTUATORII. In order to compare the performance of a normal participant and that of an ideal observer, sensitivity indices for 21 vowel pairs used in the ideal-observer case were extracted from Table 7 and Table 8. The mean of these pairs was 3.30 (std 1.16) in the last 20 runs of training and 3.32 (std 1.25) in post-training test condition. It should be noted that if the participant performed accurately in the identification experiment then the corresponding sensitivity index resulting from the 2×2 stimulus-response confusion matrix would be infinity. A saturation level of d′=4.65 was set that corresponded to the hit and false alarm rates of 99%. Performance of the ideal observer for each of the 21 vowel pairs was obtained by taking the largest d′ of the four acoustic cues and setting it to the saturation level of 4.7. Table 9 presents the largest and saturated d′ value for the 21 vowel pairs and is reproduced by merging the values shown in Tables 3, 4, 5, and 6.

Table 9. Largest d′ for vowel pairs with ideal observer

	/ae/	/ah/	/ee/	/eh/	/ih/	/oo/	/uu/
/ae/		3.1	4.7	2.8	4.6	4.7	4.5
/ah/			4.7	4.6	4.7	4.7	4.7
/ee/				4.7	1.3	4.7	4.7
/eh/					3.4	4.7	3.2
/ih/						3.3	3.3
/oo/							2.2
/uu/							

Note that all d′ values in Table 9 corresponded to features associated with ΔA_1 and ΔA_2. The average and standard deviation of sensitivity indices in Table 9 are 3.94 and 0.99, respectively. The d′ values of ideal observer (Table 9) and those in the last 20 runs of training and in post-training conditions were analyzed in a one-way ANOVA (analysis of variance) to compare the performance in the three cases, using an alpha level of 0.05. With the reported data, ANOVA failed to show that the three conditions were significantly different ($F(2,60)=2.16$, $p>0.05$).

The post-training performance of a human participant on the identification of multiple tokens of 10 vowels presented through motional and vibrational cues on the TactuatorII display averaged 42.6% with IT of 1.32 bits. This performance was comparable to that of an ideal observer operating on the acoustic cues presented through the vibrational component of the TactuatorII display. The performance scores of the participant were higher than those reported in previous studies, despite the fact that multiple vowel tokens produced by two different speakers were used. Previous research with auditory vowel perception has shown that vowel discrimination scores decreased significantly when the number of tokens per vowel was increased from one to sixteen [23]. Weisenberger et al. compared vowel identification performance on several body sites and by an optimal (based on principal component analysis) coding scheme [3]. They tested normal participants on eight vowels and with some variability in tokens (3 tokens of each vowel, either by one speaker or by three different speakers). Correct-answer feedback was provided to the participants. Identification scores were PC=61% (chance level 12.5%) when two fingers of the same hand were stimulated and IT=1.28 bits for the same speaker. For different speakers, PC=48% and IT=0.83 bits. The present study achieved PC=55.3% (chance level 10%) and IT=1.32 bits with similar experimental conditions and a larger variability in the vowel stimulus set.

In the future, we will evaluate discrimination performance of the ideal observer by using the entire corpus spoken by multiple speakers as used with the human participant (i.e, four tokens of ten pure vowels spoken by two females). We will also include speech corpus spoken by a male speaker and propose an adaptive coding scheme that updates the first and second formant bands based on the fundamental frequency of the speaker. The formants of female and children are generally greater than those for male speakers and can be derived from the voice fundamental frequency of the speaker [21]. We hope to demonstrate that the coding scheme used in the present study can be effectively applied to a wider range of speech materials, and similar performance levels can be achieved with additional participants once they have gone through a training process with the TACTUATORII device and the coding scheme.

ACKNOWLEDGEMENTS

This research was partly supported by a research grant no. R01-DC00126 from the National Institute on Deafness and Other Communication Disorders, National Institutes of Health. The first author was also partly supported by a NASA award under grant no. NCC 2-1363 and an NSF award under grant no. 0533908.

REFERENCES

[1] R. H. Gault, "Progress in Experiments of Tactual interpretation of oral speech," *Journal of Abnormal and social psychology*, vol. 19, pp. 155-159, 1924.

[2] G. Plant, J. Gnosspelius, and H. Levitt, "The Use of Tactile Supplements in Lipreading Swedish and English: A Single-Subject Study," *Journal of Speech and Hearing research*, vol. 34, pp. 172-183, 2000.

[3] J. M. Weisenberger, J. C. Craig, and G. D. Abbott, "Evaluation of a Principal-components Tactile Aid for the Hearing-impaired," *The Journal of the Acoustical society of America*, vol. 90, pp. 1944-1957, 1991.

[4] M. A. Clements, L. D. Braida, and N. I. Durlach, "Tactile communication of speech: comparison of two computer-based displays," *Journal of rehabilitation research and development*, vol. 25, pp. 25-44, 1988.

[5] C. M. Reed, W. M. Rabinowitz, N. I. Durlach, L. D. Braida, S. Conway-Fithian, and M. C. Schultz, "Research on the Tadoma method of speech communication," *The Journal of the Acoustical society of America*, vol. 77, pp. 247-257, 1985.

[6] C. M. Reed and N. I. Durlach, "Note on Information transfer rates in human communication," in *Presence*, vol. 7, 1998, pp. 509-518.

[7] H. Z. Tan and W. M. Rabinowitz, "A new multi-finger tactual display," presented at International Symposium on Haptic Interfaces for Virtual Environment and Teleoperator Systems, 1996.

[8] S. P. Eberhardt, L. E. Bernstein, D. Barac-Cikoja, D. C. Coulter, and J. Jordan, "Inducing dynamic haptic perception by the hand: system description and some results," presented at American Society of Mechanical Engineers: Dynamic Systems and Control, 1994.

[9] I. R. Summers, "Tactile aids for the hearing impaired." London: Whurr Publishers Limited, 1992.

[10] S. J. Bolanowski, G. A. Gescheider, R. T. Verrillo, and C. M. Checkosky, "Four channels mediate the mechanical aspects of touch," *The Journal of the Acoustical society of America*, vol. 84, pp. 1680-1694, 1988.

[11] R. W. Cholewiak and A. A. Collins, "Sensory and physiological bases of touch," in *The Psychology of Touch*, M. A. Heller and W. R. Schiff, Eds. Hillsdale, N. J.: Lawrence Erlbaum Associates, 1991, pp. 23-60.

[12] A. Israr, H. Z. Tan, and C. M. Reed, "Frequency and amplitude discrimination along the kinesthetic-cutaneous continuum in the presence of masking stimuli," *The Journal of the Acoustical society of America*, vol. 120, pp. 2789-2800, 2006.

[13] H. Z. Tan, N. I. Durlach, C. M. Reed, and W. M. Rabinowitz, "Information transmission with a multifinger tactual display," *Perception and Psychophysics*, vol. 61, pp. 993-1008, 1999.

[14] A. Israr, P. H. Meckl, and H. Z. Tan, "A two DOF controller for a multi-finger tactual display using a loop-shaping technique," presented at 2004 ASME International Mechanical Engineering Congress and Exposition (IMECE04), Anahiem, CA, 2004.

[15] N. A. Macmillan and C. D. Creelman, *Detection Theory: A User's Guide*, Second ed. New York: Lawrence Erlbaum Associates, 2004.

[16] H. Yuan, "Tactual Display of Consonant Voicing to Supplement Lipreading," in *Department of Electrical Engineering and Computer Science*. Cambridge: Massachusetts Institute of Technology, 2003, pp. 257.

[17] K. W. Grant, L. H. Ardell, P. K. Kuhl, and D. W. Sparks, "The Contribution of Fundamental Frequency, Amplitude Envelope, and Voicing Duration Cues to Speechreading in Normal-Hearing Subjects," *The Journal of the Acoustical society of America*, vol. 77, pp. 671-677, 1985.

[18] R. T. Verrillo and G. A. Gescheider, "Perception via the sense of touch," in *Tactile Aids for the Hearing Impaired*, I. R. Summers, Ed. London: Whurr Publishers, 1992, pp. 1-36.

[19] R. Jakobson, C. G. M. Fant, and M. Halle, *Preliminaries to Speech Analysis*. Cambridge: MIT Press, 1963.

[20] S. A. Zahorian and M. Rothenberg, "Principal-components Analysis for Low-redundancy Encoding of Speech," *The Journal of the Acoustical society of America*, vol. 69, pp. 832-845, 1981.

[21] J. Hillenbrand, L. A. Getty, M. J. Clark, and K. Wheeler, "Acoustic Characteristics of American English Vowels," *The Journal of the Acoustical society of America*, vol. 97, pp. 3099-3111, 1995.

[22] W. R. Garner, *Uncertainty and Structure as Psychological Concepts*. New York, USA: John Wiley & Sons, Inc., 1962.

[23] R. M. Uchanski and L. D. Braida, "Effects of token variability on our ability to distinguish between vowels," *Perception and Psychophysics*, vol. 60, pp. 533-543, 1998.