Speaking Haptically: from Phonemes to Phrases with a Mobile Haptic Communication System

Maurício Fontana de Vargas, David Marino, Antoine Weill--Duflos, *Graduate Student Member, IEEE*, Jeremy R. Cooperstock, *Member, IEEE*

***Abstract*—We present three studies involving WhatsHap, a mobile system designed to deliver speech as vibrations on the forearm with minimal hardware demands and practice time. After only 4.2 h of training on a 24-haptic phoneme vocabulary and on how to combine these to form words, participants were able to generalize their phoneme identification skills to the understanding of untrained English words, correctly identifying 65% of words in phrases rendered with a user-controlled interval between words, and up to 59% with a fixed interval. Ultimately, participants were able to complete 88% of simple communicative tasks that elicited spontaneous speech and semi-structured bidirectional conversation using the apparatus. We conclude by providing insights as to how such a system may ultimately be used for communication under more natural conditions.**

***Index Terms*—phonemic coding, tactile speech communication, language acquisition, speech-to-haptic**

I. Introduction

THE tactile sense offers a useful channel for communication, especially in contexts where the visual or auditory modalities are occupied with other tasks, e.g., while driving or performing surgery, or otherwise compromised, e.g., for individuals with visual or auditory difficulties. Haptic communication frequently employs a small vocabulary of haptic icons (or "tactons" [1]), tailored to application-specific problem domains such as navigation guidance [2], mobile phone alerts [3], or numeric information delivery [4].

Another class of haptic communication is concerned with transmission of a spoken language, first popularized by the Tadoma method, in which deaf-blind users receive speech by placing their hand on the talker's face. With Tadoma, trained individuals were able to achieve a recognition accuracy of 80% for keywords in conversational sentences [5].

Initial attempts to create tactile speech aids, beginning in the 1970s, used vocoders, which converted speech directly to vibrotactile stimulation based on acoustic properties of the input speech signal [6]. This technique is able to transmit speech haptically at the same rate as the spoken input. However, the complexity of the vibration patterns associated with arbitrary input signals, including variations in pitch and speech rate, imposes extensive training requirements. For instance, hearing participants required 55 hours of training to achieve 80% accuracy on a 150-word set [7], and deaf children trained for 48 weeks (230 hours) to reach 84% accuracy on a set with 152 words [8]. Furthermore, such vocoder-based approaches do not generalize well to the recognition of untrained words.

Modern research on haptic speech communication has focused on delivering speech in terms of small discrete units such as letters or phonemes. This is in part motivated by the proliferation of text-based messaging and recent advancements in speech-to-text technology. Examples of letter-to-haptic mapping include the use of simplistic unistroke patterns resembling manual writing, rendered through four actuators worn on the wrist [9], and abstract overlapping spatio-temporal stimulations rendered through actuators worn on the hand [10], [11]. Phoneme-based mapping offers the advantage of rendering words using shorter stimuli, given that any English word is composed by a number of phonemes smaller than the number of letters. The trade-off is a larger number of basic discrete units needed to be encoded by the system, since there are 44 English phonemes and 26 letters in the alphabet. This may impose non-trivial hardware requirements, such as in Reed et al. [12], in which 24 actuators were used to deliver a set of 39 English phonemes encoded as vibrations. These actuators are mounted on a cumbersome gauntlet that must be carefully positioned and calibrated on the forearm, which restricts mobility. In addition, the hardware needed to drive the 24 actuators cannot be easily miniaturized and imposes high energy requirements, further hampering the system's mobility. A simpler apparatus capable of encoding 23 English phonemes using a multi-sensory (radial squeeze, lateral skin stretch, vibration) set of stimuli was introduced by Dunkelberger et al. [13]. After 100 minutes of training on a set of 150 words, participants were able to correctly identify 87% of words from a 12-item list, with a self-paced rate of phoneme rendering.

The aforementioned work demonstrated the feasibility of delivering speech through the sense of touch, relying on a discrete mapping. However, we are not aware of any instances in which such an encoding has been employed under more realistic scenarios, such as receiving untrained phrases constructed with an extensive and complex vocabulary, or in bidirectional communication, where two interlocutors are able to carry out an unstructured "haptic conversation".

This article reports on three studies involving "WhatsHap", a mobile system designed to deliver text or speech as vibrations on the forearm with minimal hardware demands and training time. Naive participants engaged in self-training activities for a total of 4.2 hours, spread over a 14-month

Maurício Fontana de Vargas is with the School of Information Studies and the Centre for Interdisciplinary Research in Music Media and Technology, McGill University, Montreal, Canada

David Marino, Antoine Weill–Duflos, and Jeremy R. Cooperstock are with the Department of Electrical and Computer Engineering and the Centre for Interdisciplinary Research in Music Media and Technology, McGill University, Montreal, Canada.

Manuscript received April 19, 2005; revised August 26, 2015.

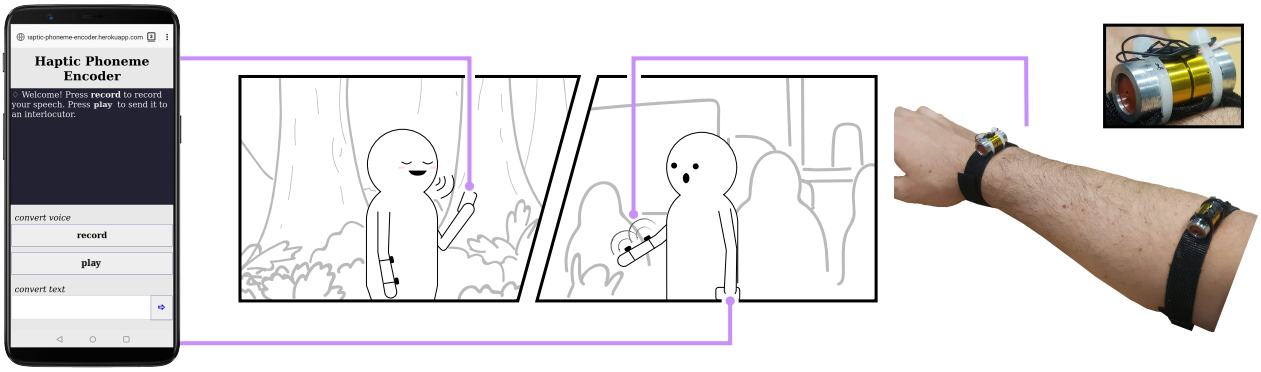

Fig. 1: WhatsHap is a messaging app that facilitates haptic conversation between two or more interlocutors. A user may speak or input text into the app, and the constituent phonemes of their message are sent as a sequence of patterned vibrations to another user's arm. Users may reply in turn, enabling real time haptic conversation.

period, to learn a subset (24) of English phonemes rendered haptically, advancing to words, phrases, and ultimately, simple communicative tasks where the conversation partner's speech was delivered solely by vibrotactile actuation. A summary of these studies and their respective goals is provided below. For details on the number of phonemes, words, and phrases used in the studies, please refer to Table I.

Study 1: fourteen participants were introduced to the apparatus and performed 100 minutes of self-training spread across four consecutive days. Activities were designed to teach phonemes progressively, allowing participants to focus on learning a small subset of phonemes at a time, while reviewing and reinforcing the phonemes previously learned. Participants were also trained on how to combine phonemes to form words and practiced word identification with a 150-word list. Testing activities evaluated the word-recognition ability both from a multiple-choice list and without any visual cuing of possible choices. This study was reported previously [14].

Study 2: seven participants from the first study returned twelve months later to perform an additional 150 minutes of training in five days. After reviewing the haptic phoneme encoding learned in the first study, participants practiced and were evaluated on word identification with larger and more complex word sets (a total of 514 words), and finally on phrase identification with an inter-word interval (IWI) that was either pre-determined (3 s) or user-controlled. Performance analysis included the impact on word-level comprehension of words containing phonemes not covered by our initial 24 phoneme encoding (36% of word occurrences) and how contextual information obtained from identified words in a phrase impacts both word- and phrase-level comprehension. The findings from this study constitute the major contribution of this article.

Study 3: the three best performing participants from Study 2 were recruited two months later to engage in conversation tasks with two naive conversation partners (CP) with experience in speech science. The study involved the previous participants receiving spoken messages from the CPs as haptic stimuli, and responding through text. The conversational tasks focused on achieving a certain goal (e.g., scheduling a time to watch a movie), rather than a purely linguistic outcome. The main goal of this study was to understand *how* users input linguistic content to the system and the user *experience of understanding* a message encoded as haptic phonemes.

II. WhatsHap Apparatus

WhatsHap consists of two vibrotactile transducers (Haptuator Original, Tactile Labs, Montreal, Model no. TL002-14-A) [15] attached to armbands and connected to a smartphone or microcomputer. The app renders vibotactile signals through its standard audio output. One of the armbands is worn near the wrist, and the other close to the elbow, such that the actuators are in contact with the dorsal side of the user's forearm, as illustrated in Fig. 1. Vibration patterns representing individual English phonemes are stored as standard stereo audio files. Words are rendered as a sequence of haptic phonemes with a 1 s inter-phoneme interval (IPI), and phrases with a 3 s IWI.

The app conveys text or speech haptically through a web-based messaging software, using the Google Speech-to-Text API[1] to convert the utterances to text. It then obtains a phonemic representation of the words in North American English with the CMU Pronouncing Dictionary[2] and finally, broadcasts haptic messages to interlocutors using the apparatus.

A. Phoneme-to-vibrations mapping

Our design supports 24 of the approximately 44 English phonemes: 15 consonants (/p, b, t, d, k, ɡ, f, v, ð, s, z, m, n, l, ɹ/), 5 vowels (/i, ɛ, ʌ, u, ɑ/), and 4 diphthongs (/eɪ, ɑɪ, ɑʊ, oʊ/). These were selected based on frequency of use in casual conversation [16] and simplification of inter-phoneme similarity in which similar-sounding phonemes are replaced by a common haptic stimulus, where feasible. For example, the word *book* (/bʊk/) is rendered by replacing the missing phoneme /ʊ/ with a similar one, i.e., /u/ (e.g., as in *boot*). This approach is motivated by the fact that non-native English speakers may successfully communicate despite not being completely perceptually sensitive to its phonemic contrasts [17] [18], and that native listeners are able to integrate top-down and bottom-up processing to comprehend words despite superficial phonological errors [19]. This allows the rendering

[1] https://cloud.google.com/speech-to-text/docs/
[2] http://www.speech.cs.cmu.edu/cgi-bin/cmudict

of a larger set of words while keeping the simple hardware design and reducing the burden on users when learning the mapping and mentally decoding the vibrations. Ten phonemes are rendered with their substitutes: æ→ɛ, j→i, ɪ→i, ɝ→ʌɹ, ʊ→u, ɔ→ɑ, w→u, ŋ→ng, ɔɪ→ɑɪ, θ→f.

The haptic stimuli were created from the audio waveforms of a phoneme's exemplary phone, enhancing salient physical characteristics during normal speech, e.g., the place in the vocal tract where the sound is articulated, and the vibration intensity of the vocal chords. This strategy provides a more natural mapping that facilitates learning and allows users to generalize to recognizing novel haptic words. Fig. 2 illustrates the rendering strategy. The haptic stimuli design is summarized below. Further details are provided in reference [14].

1) Audio of isolated consonants was obtained from recordings of a native English speaker (jbdowse.com/ipa); audio from vowels and diphtongs was obtained from computer synthesized speech using Praat (praat.org).
2) The raw audio signal was processed to enhance distinct features inherent to how speech organs are involved when producing the sound. For example, the high-frequency, turbulent sound of fricative phonemes (/f, v, s, z/)–caused by the air going through a narrow gap between the lips or teeth–was emphasized by a high pass filter. The strong, short puff of air characteristic of the plosives (/p, t, k/) was characterized by changes in gain. Vowels present the same manner of articulation and thus were not enhanced in this manner. To improve discrimination between vowels and consonants, a unique fade-in and fade-out effect was applied to all vowels.
3) The interaural level difference (ILD) of the two channels was manipulated to create a spatial (wrist/elbow) panning indicative of the phoneme's place of articulation within the vocal tract. Phonemes produced towards the front of the mouth (e.g., /p, f, l, i/) are mapped to distal region of the forearm, while phonemes produced towards the back of the vocal tract (e.g., /k, ɹ, u/) are rendered in the proximal region. The articulation movement characteristic of diphthongs is mimicked by linearly varying the ILD between the values of the vowels composing the diphthong, resulting in the perceptual illusion that the vibration location moves.
4) Finally, audible frequencies not detected by skin receptors were removed with a low-pass filter with cutoff frequency of 700 Hz.

III. Study 1 – Phonemes and Words

This study focused on delivering basic training on haptic phonemes to naive participants, and evaluating the feasibility of using the adopted encoding strategy to render full words. To facilitate comparison of results, our experimental design, including training and testing activities, was inspired by previous work on the topic by Dunkelberger et al. [13].

A. Participants

We recruited fourteen participants (8 male, 22-43 years of age, $\bar{x} = 29, \sigma = 6$) through university email lists. Only one participant (P4) was a native English speaker, one (P11) reported being non-fluent, and the remaining were fluent. Five participants (P1, P7, P3, P5, P14) reported having some prior knowledge of phonetics, and six had received music training.

All participants provided informed consent of the experiment protocol, following Research Ethics Board guidelines, and received compensation of CAD $60.

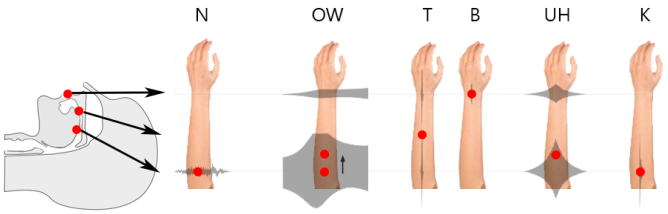

Fig. 2: Mapping between the vocal tract location where the phonemes composing the word "notebook" are articulated and the approximate position in the forearm where the corresponding vibration pattern is felt. Stimuli waveforms are plotted in the background, with the left channel underneath the wrist and the right channel underneath the elbow.

B. Experimental Environment

The study was held in a laboratory setting. All training and testing activities were self-guided and delivered through a desktop computer. Participants sat in front of the computer and interacted with the software using a mouse, wearing the arm bands on the right arm and over-ear headphones with pink noise played at a comfortable level to block out all residual sound produced by the vibrotactile transducers.

C. Training and testing protocol

Participants performed 100 min of self-training, spread non-uniformly over four consecutive days. Training consisted of i) a free-play panel, in which participants could experiment with all phonemes being learned, ii) phoneme identification quizzes, and iii) word construction quizzes, designed to teach participants how to combine haptic phonemes to form words.

On days 2–4, participants first performed a pre-test on all phonemes learned so far. They ended with a test of recognition ability on random words rendered with participant-controlled inter-phoneme interval (IPI) (Post-Tests and Final Test) and with a fixed IPI of 1 s (Sequential Post-Test), supported by a 12-option answer list. On the last day, participants typed their answers to the Open Answer Final Test (OAFT), in which words were rendered with an IPI of 1 s. This test consisted of 25 words from the 150-word training set and 25 completely new words, ranging from two to six phonemes ($\mu = 3.1$).

D. Results and Discussion

Excluding one non-compliant participant, the average phoneme identification score in the final day pre-test was 74.7%. We found that most errors involved fricatives and vowels (57.5% and 65.2% average accuracy, respectively), while the other groups presented higher accuracy rates: 92.0%

for plosives, 85.3% for nasals and approximants, and 71.8% for diphtongs. In the Open Answer Final Test (OAFT), in which phonemes appeared forming a word and were rendered with a fixed IPI of 1 s, average phoneme identification accuracy was 68.0%.

Final Test (FT) word accuracy, where participants advanced through the phonemes composing a word at their own pace was 94.4%. This compares favorably to the 86.6% observed in previous work that followed a similar training and testing protocol [13]. In the OAFT, in which a novel, challenging condition was implemented (i.e., words rendered with a fixed IPI and without presentation of a list of response options), average word accuracy score was 45%.

When comparing word identification accuracy between participant-controlled phoneme advance (Post-Tests) and fixed IPI condition (Seq. Post-Test), we found only a slight decrease (7% on Day 3, not statistically significant on a Student's t-test) for words rendered with a fixed IPI. This suggests that the lack of response clues was the main factor to the decrease in accuracy observed on OAFT in comparison with FT.

Difficulties in identifying particular phonemes may also have affected OAFT performance. Given our mapping strategy, it was expected that participant mistakes would typically occur between similar-sounding phonemes (i.e., those having a similar place or manner of articulation), which would have only minor impact on word comprehension.

This was indeed observed for plosives, vowels, nasals and approximants. However some fricatives and vowels were often mistaken for very different phonemes, (e.g., /z/↔/u/; /s/↔/a/; /ɛ/↔/v/), transforming the sequence of phonemes into a meaningless, nonsensical sound, preventing participants from understanding the rendered word.

Considering the limited training time relative to the significant task of language acquisition, we are encouraged by the observed scores, which demonstrate the feasibility of learning to interpret words rendered as a sequence of vibration patterns, constructed from the audio of the constituent phonemes. Participants were picking words from a list of possible responses during the Final Test, which is an easier task than freely entering responses. Nevertheless, the high accuracy scores on phoneme retention obtained on pre-tests, and especially on OAFT, strongly indicate that participants identified words based on their constituent phonemes rather than memorizing chunks of stimuli mapping entire words. Readers interested in further details may find these in reference [14].

IV. Study 2 – Phrases

A. Participants

We invited all thirteen participants who were compliant with the instructions from the first study to participate in this follow-up twelve months later. Seven participants (P2, P4, P5, P6, P7, P8, P13) returned. Their average phoneme-recognition ($\mu = 73\%, \sigma = 17\%$) and word-identification ($\mu = 53\%, \sigma = 14\%$) scores in the last day of Study 1 are reasonably similar to the scores of the entire set of participants from Study 1 ($\mu = 75\%, \sigma = 14\%$ on phonemes and $\mu = 45\%, \sigma = 16\%$ on words).

All participants provided informed consent, following Research Ethics Board guidelines, and received compensation of CAD $100 after completing the experiment.

B. Training and testing protocol

Participants performed 150 min of self-training, spread across five consecutive days, as illustrated in Fig. 3. The study was divided into four blocks, progressing from a simple review (Block 1) to understanding of entire phrases rendered haptically (Block 4). Each block began and ended with a test activity, with several rounds of training in between. The experimental environment was the same as in Study 1.

Our protocol considers findings from the first study showing that individual phoneme identification ability led to successful word comprehension. Therefore, phoneme identification was reinforced in the beginning of each session, constituting half of the training time in Block 1 and one third of the training time in Block 2. In addition, participants could switch to the Free-Play panel at any time if they were encountering difficulties on more advanced tasks on Blocks 1 and 2.

Block 1 was performed on the first day and served as a review session containing only activities on phoneme and word identification using the same 150-word dataset from Study 1. New training activities were introduced so that participants could gain more experience with words rendered with a fixed IPI and responding in an open-answer format, in order to bridge the gap in difficulty between training and testing phases informally reported by participants in Study 1.

On the second day, participants performed Block 2, which continued reinforcing individual phonemes and familiarizing participants with full-word rendering and open-answer quizzes. The word identification score obtained in the post-test was used as a benchmark to determine whether participants would advance to the more challenging activities of Block 3, involving phrases, or would repeat Block 2 the following day. Participants with an average word-identification score lower than 60% and a normalized phonological edit distance greater than 1.0[3] repeated Block 2 the following day. This combination of conditions was adopted so participants with accuracy scores slightly lower than the desired 60% could still advance to phrase training, as long as most of their errors were very close to the correct words. Regardless of their performance on Block 2, all participants were permitted after their third attempt to advance to Block 3.

Block 3 was dedicated to practicing phrase recognition at a user-determined pace, i.e., participants controlled when the next word in the phrase was rendered, and with phrases constructed from a more advanced word set. The protocol was designed such that all participants participated in this block exactly once so that their performance could be compared.

Block 4 was intended as an extra challenge to investigate participants' ability to understand entire phrases, constructed from an even more challenging vocabulary with a fixed IWI

[3]The phonological edit distance is the number of operations needed to transform one string to another, weighted by distinctive features [20], [21]. Lower scores correspond to closer words. Values are normalized by the number of phonemes in the word. Some examples: a-eye= 4.3, soon-son= 1.7, snow-rain=6.9, won't-work=4.3

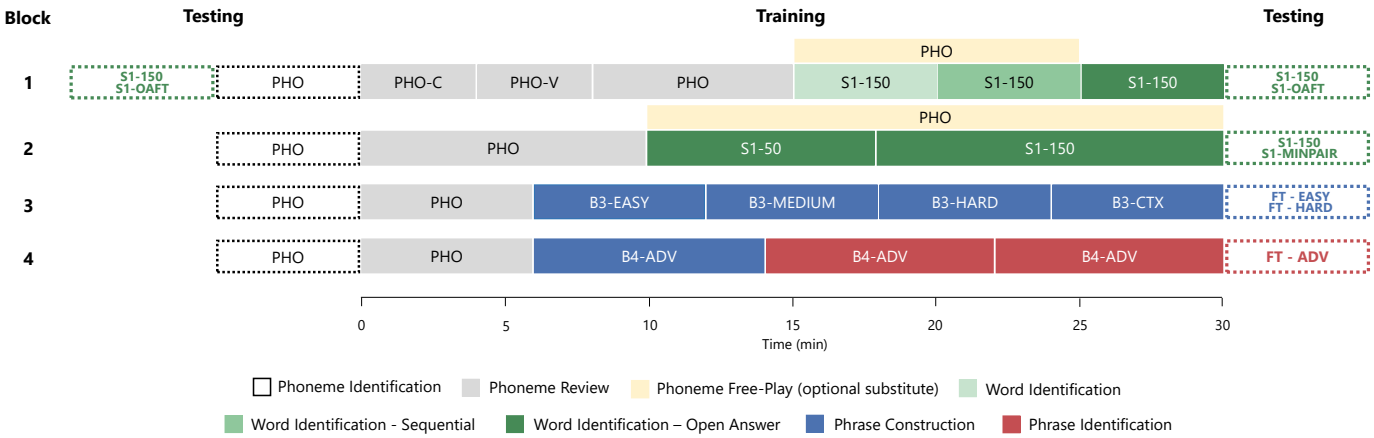

Fig. 3: Training and testing protocol for Study 2. Training (filled-in boxes) is time-limited, while test activities (dash-outline boxes) have a fixed number of trials. The data sets used are indicated in each activity. S1-150: words from Study 1; S1-50: shortest words from S1-150; PHO-C: consonant phonemes; PHO-V: vowel phonemes; S1-OAFT: untrained words from Study 1 Open Answer Final Test; S1-MINPAIR: words differing by a single phoneme from words in S1-150; CTX: phrases related to a context, e.g., shopping for clothes; ADV: advanced phrases.

of 3 s. Only participants who had passed Block 2 within two attempts performed this block. Since Block 3 was performed only once, participants who successfully completed Block 2 on the first attempt performed Block 4 twice.

Following the last session, participants completed a short questionnaire regarding their experience using the apparatus during training and testing.

We now describe each of the training and testing activities adopted in this study, and provide further details of the word and phrase sets used, as summarized in Table I.

C. Training Activities

We extended Study 1 activities to include practice on word formation with a fixed inter-phoneme interval (IPI), in addition to the training of entire haptic phrases, as summarized below. All training quizzes are time-limited and provide correct-answer-feedback, delivered visually and haptically.

Phoneme Free-Play: participants can experiment with all phonemes being learned. A diagram with the vocal tract when uttering the phoneme and a visual representation of the vibration stimulus are also displayed.

Phoneme Review with Pre-Test: participants perform a phoneme identification Pre-Test, and are able to review all phonemes through the Free-play panel, in which scores are displayed beside each phoneme. Once participants feel confident, they advance to a phoneme identification quiz.

Word Identification Quiz: a random word is selected from the same set used in Study 1 (S1-150) and is rendered with an user-controlled IPI. Participants are requested to identify the word from a multiple-choice list containing 12 random words.

Word Identification – Sequential: same as Word Identification Quiz but with a IPI of 1 s.

Word Identification – Open Answer: same as the previous activity, but without the multiple-choice list. Participants type their answers in a single text box.

Phrase Construction: a randomly selected phrase from our phrase set[4] is rendered, one word at a time. Participants control advancement. They can type their answer at any time, each word in a specific text box, and submit the entire response after the last word is rendered. The phrase sets used in this activity increase in difficulty over time in terms of the portion of new words (not in S1-150) constituting each phrase, and include vocabulary rendered with substitute phonemes (see Table I for details). Participants were not explicitly informed whether a word included a substitute phoneme.

Phrase Identification: a longer random phrase ($\mu = 5.1$ words/phrase) than used in the previous sets is rendered with a fixed IWI of 3 s. Participants type their answer in a single text box. Most of the phrases (88.7%) include at least one word containing a substitute phoneme. Individual words rendered with a substitute phoneme account for 42.3% of word occurrences within the set of phrases. As in the previous activity, participants were not explicitly informed whether words were being rendered with a substitute phoneme.

D. Testing activities

These activities are performed in the same manner as their equivalently named training counterparts, but with a fixed number of trials[5] and without providing correct-answer feedback. Thus, they do not count towards training time.

Block 1 – Word Identification Open Answer: Twenty words are rendered: ten from S1-150, and another ten from the untrained set S1-OAFT. This is the same test as the Study 1 Open Answer Final Test, differing only in number of trials.

Block 2 – Word Identification Open Answer: a total of twenty words are rendered: ten from S1-150 and another ten from S1-MINPAIR, a 101-word set containing S1-150 minimal pairs, which differ only by a single phoneme (e.g.,

[4]Available at srl.mcgill.ca/toh/hapticspeech

[5]Participants encountering major difficulties were asked to stop after 15 min, which was the case for only one participant (P6).

TABLE I: Summary of word and phrase sets used during training and testing, available at srl.mcgill.ca/toh/hapticspeech

Set Name	Phrases	Words	Phonemes (μ, σ)	Words/Phr. (μ, σ)	Subs.[a] (%)
S1-50	-	50	2.6, 0.7	-	0.0
S1-150	-	150	3.1, 1.0	-	0.0
B3-EASY[b]	38	86	3.2, 1.2	3.3, 0.8	13.5
B3-MEDIUM[c]	52	122	3.2, 1.1	4.0, 1.1	28.0
B3-HARD[d]	25	67	3.3, 1.3	4.0, 1.1	37.8
B3-CTX[e]	35	84	3.7, 1.4	3.3, 1.0	46.1
B4-ADV	159	368	3.7, 1.5	5.1, 1.2	42.3
S1-OAFT	-	50	3.1, 0.9	-	0
S1-MINPAIR[f]	-	101	3.0, 0.6	-	0
FT-EASY[g]	12	34	2.9, 1.1	3.5, 0.5	35.7
FT-HARD[h]	12	35	3.4, 1.3	3.3, 0.5	35.9
FT-ADV	25	99	3.3, 1.2	5.6, 1.3	50.0
ALL	274	514	3.7, 1.4	4.4, 1.5	40.9[i]

[a] Word occurrences rendered with at least one substitute phoneme.
[b] 0–33% [c] 34–66% [d] 67–100% new words
[e] Phrases related to a specific context (e.g., shopping for clothes)
[f] Words differing by only one phoneme from words in S1-150
[g] 0–33% [h] 67-100% untrained words
[i] 8.4 with more than one substitute phoneme

make–lake). The score achieved in this test determines whether a participant should advance to phrase training.

Block 3 Final Test – Phrase Construction: ten untrained phrases are rendered in total: five from FT-EASY, with phrases having at most 33% of untrained words, and five from FT-HARD, in which phrases contain at least 66% completely new words. This test investigated the extent to which additional linguistic information, i.e., the surrounding words in the phrase, could be leveraged to improve individual word recognition.

Block 4 Advanced Test – Phrase Identification: ten untrained phrases selected from FT-ADV are rendered with a predetermined interval between words. Sentences are longer than in the previous tests ($\mu = 5.6$ words/phrase), and all include at least one word rendered with a substitute phoneme.

E. Results

1) Remembering and identifying phonemes: The phoneme identification accuracy scores, defined as the number of correctly identified phonemes divided by the number of trials on the daily pre-tests are shown in Fig. 4. The scores from the Study 1 Open Answer Final Test are also plotted for comparison. On average, participants remembered 22.2% of phonemes after one year without using the apparatus. With an additional 30 min of training (15 min dedicated exclusively to phonemes), participants were able to reach a level of performance similar to that from Study 1 (64.7% vs. 73.0%). Their average score from Study 1 was surpassed after 60 min of additional training and continued to increase as more training was performed. After the review session (Block 1), scores increased at a linear ($R^2 = 0.98$) rate of 13.4 %/h or the equivalent of 3.2 new phonemes per hour.

2) Understanding words: Participants were not at first able to recognize words after one year since their last contact with

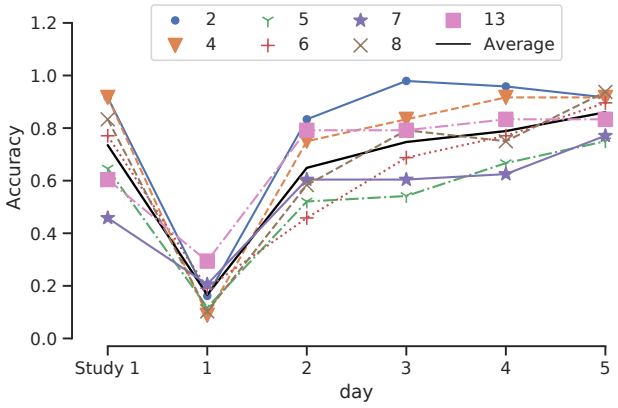

Fig. 4: Pre-test phoneme identification accuracy scores. Average scores across all participants are shown as the black line.

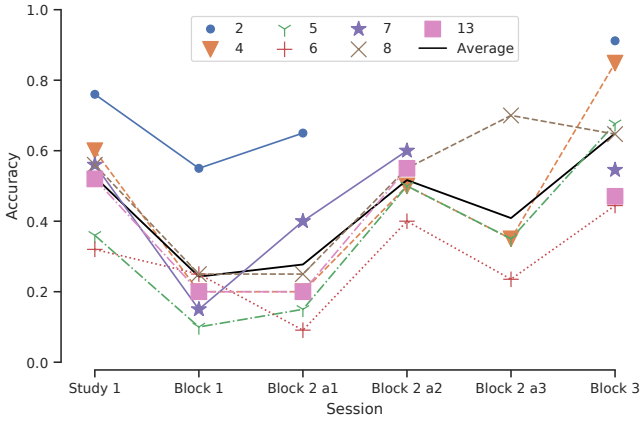

Fig. 5: Post-test word identification accuracy. Average scores consider only those participants performing the particular session. Blocks 2 a1–a3 represent the three possible attempts to achieve the minimum score for advancing to Block 3.

the apparatus (Block 1 Word Id. Pre-test $\mu = 2\%$). Fig. 5 shows how word identification accuracy scores (i.e., number of correctly identified words divided by the number of quiz trials) evolved throughout training. On average, participants presented little progress in the first 60 min (Block 2 a1) (28.4%), but were able to reach 53.6% after one more session (Block 2 a2), similar to the accuracy obtained in the first study (52.3%). One participant (P2) reached the 60% threshold on word identification accuracy to advance to phrase training in the first Block 2 attempt. Two more participants advanced to Block 3 after one additional training day (P7 with 60% accuracy and P13 with a normalized phonological edit distance of 0.6). The remaining four participants performed Block 2 for one last additional session but only P8 demonstrated improved accuracy (55.0% to 70.0%). We cannot conclude whether these participants performed worse due to external factors, such as fatigue or stress, or because they had reached the limit of their ability after two attempts. Additional days of training with the same block would be required to answer this question.

To better understand the severity of participants' mistakes,

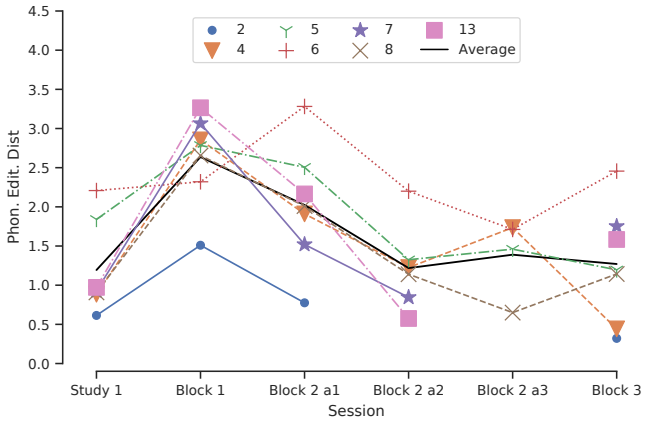

Fig. 6: Normalized phonological edit distance on post-tests.

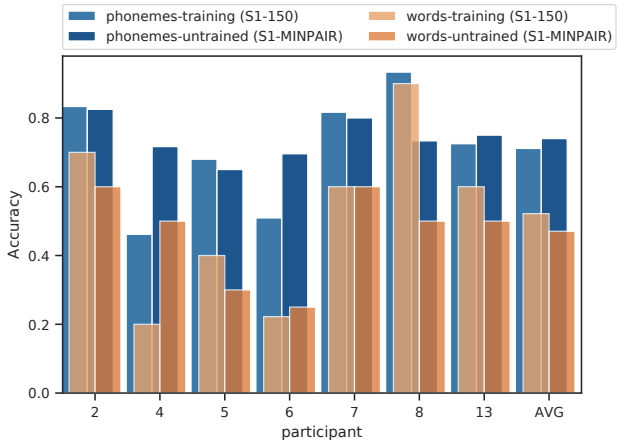

Fig. 7: Word and phoneme identification accuracy scores for words in the training set (S1-150) and untrained minimal pairs (S1-MINPAIR) for participants' last session of Block 2.

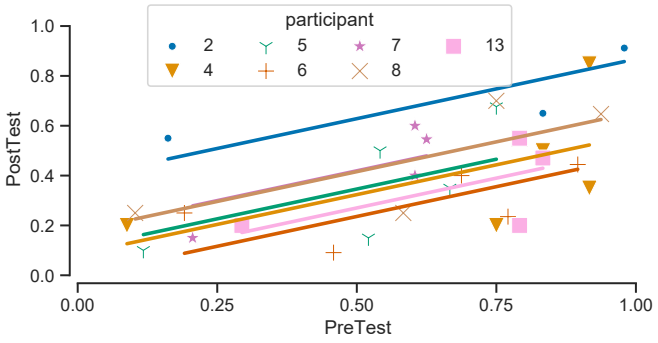

Fig. 8: Repeated measures correlation of the accuracy scores between phoneme identification (pre-tests) and word identification (post-tests). Data are plotted for each participant to demonstrate that the general fit applies individually.

we also calculated the normalized phonological edit distance throughout the training sessions, as presented in Fig. 6. Although participants did not improve significantly in terms of word identification accuracy after the first Block 2 training session (Block 2 a1) (4.2%), the 0.6 decrease (from 2.6 to 2.0) in phonological edit distance in the same time frame indicates that they were comprehending a larger number of phonemes and getting closer to the correct words. The average phonological distance continued to decrease with an additional 30 min of training (Block 2 a2), reaching a level that permitted understanding of entire words. This can be observed in the large increase (28% to 51%) in word identification accuracy score from the plateau of Block 2 a1 to a2.

We wished to determine whether participants would rely on the constituent haptic phonemes or would attempt to memorize and associate the entire stimulus sequence as a word. As shown in Fig. 7, the similar accuracy of word identification on the training set (S1-150) and untrained words (S1-MINPAIR), achieved during the last session on Block 2, suggests the former. A statistical test would not be meaningful here given the limited number of participants in the sample. We also calculated the correlation between phoneme (pre-tests) and word identification scores (post-tests). A repeated measures correlation test [22] attained a $p_{value} < 0.01$, confirming significant correlation with a strong association (rmcorr coefficient of 0.68 and achieved power of 0.97). Fig. 8 shows the linear fit obtained by aggregating the data across all participants, plotted for each participant's data.

3) Words rendered with a substitute phoneme: We compared participants' performance on trials of "mispronounced" words (i.e., incorporating one or more phonemes from a substitute set, as a strategy to cover a larger number of words without additional haptic symbols) against trials of words composed entirely of exact phonemes from the covered subset during the Block 3 post-test. Performance on the two groups was of a similar level of accuracy ($\mu = 62.5\%, \sigma = 19.5\%$ for the words with one or more substitute phonemes vs. $\mu = 66.5\%$, $\sigma = 17.5\%$ for the words with all-exact phonemes). However, the difference in the average phonological edit distance per word ($\mu = 6.1, \sigma = 3.4$ vs. $\mu = 3.6$, $\sigma = 2.3$) indicates that the use of a substitute phoneme may aggravate the participants' errors when decoding the word.

4) Understanding words within a phrase: The word identification accuracy scores from Block 3 post-test, in which words appeared within the context of a phrase rather than individually, are plotted in Fig. 9.

Participants correctly identified an average of 65% of the rendered words. The normalized phonological edit distance ($\mu = 1.3, \sigma = 0.7$) indicates the severity of their mistakes: these were typically of the form of understanding *slow* as *snow* (or *goat* as *gate*) for every rendered word. Comparing the accuracy scores for words appearing in phrases of Block 3 against the highest scores achieved during pure word identification tasks in Block 2 (Fig. 5), four participants (P4, P2, P5, P6) demonstrated improvement (70%, 40%, 35%, 11%, respectively), while the remaining three (P7, P8, P13) demonstrated a relative decrease in performance (-9%, -8%, -14%). On average, accuracy increased from 56% to 65%. No significant difference was found by a Friedman test ($p > 0.1$) between the easy and hard phrase sets in terms of word accuracy (65% vs. 66%), nor normalized phonological edit distance ($\mu = 1.2$ vs. $\mu = 1.3$, $\sigma = 0.6$).

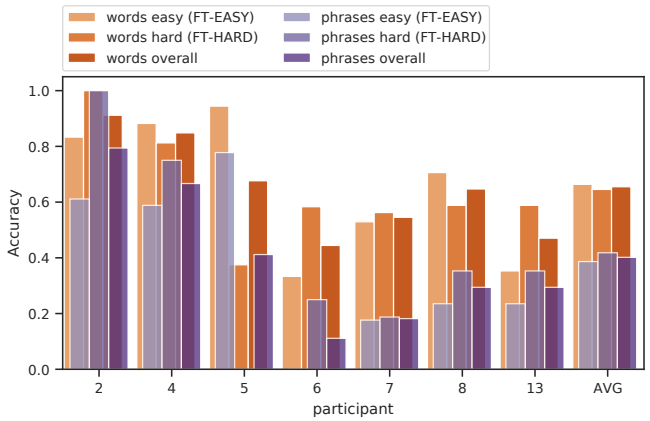

Fig. 9: Accuracy scores in terms of phrases and words correctly identified on Block 3 – post-test.

5) Understanding Phrases: Fig. 9 presents accuracy scores on phrases and their constituent words in Block 3 – Phrase Construction post-test. On average, participants recognized the entire phrase in 42% of trials, with similar scores between easy and hard phrases sets, indicating that participants were able to generalize their haptic phoneme knowledge to the understanding of untrained vocabulary.

6) Advanced Phrases: One participant (P2) reached Block 4 by the fourth experimental session, and thus performed this block twice, correctly identifying 47.3% and 58.5% of the words, with normalized phonological edit distance of 2.5 and 1.7, indicative of the increased difficulty of recognition with fixed inter-word intervals. Examples of some of the best trials include "please tell me if you feel any pain", in which P2 only omitted "any", the successful "get into the room" and "when are you going?". Two other participants (P7 and P13) performed Block 4 once, attaining word identification accuracy of 33.3% and 29.6%, and normalized phonological edit distance of 2.8 and 3.0, respectively.

7) Post-Questionnaires: Fig. 10 presents participants' responses regarding their experience with the apparatus and mapping strategy, measured on a 5-point Likert scale.

F. Discussion

Participants could still remember some of the haptic phonemes one year after last using the apparatus. Nevertheless, this residual knowledge did not lead to a faster learning pace in the second study, as demonstrated by the phoneme and word identification accuracy, as well as the phonological edit distance scores throughout the sessions.

The similarity of scores obtained on untrained and trained words, in addition to the high correlation between phoneme and word identification accuracy scores, indicates that participants identified words based on their constituent phonemes, instead of memorizing and associating chunks of stimuli with words. This was also observed in phrase understanding results, for which participants demonstrated similar performance, regardless of whether the phrases were constituted predominantly of words from the trained or untrained vocabulary. Indeed, this was confirmed by the responses to our post-questionnaire, shown in Figure 10.

Substituting similar-sounding phonemes with a common haptic stimulus allowed us to cover 34 phonemes with only 24 distinct haptic representations, consequently minimizing hardware demands. Significantly, this strategy allows to convey 95% of phoneme occurrences in conversational English [16], compared to 71% when using only the original subset of 24 phonemes. The drawback is that minimal pairs involving such potentially substitute phonemes (e.g., man–men, eat–it, peel–pill) can only be correctly identified with the support of contextual information such as the other words in the phrase.

Direct comparison of accuracy scores with other literature is unfortunately not straightforward due to significant differences between experimental conditions, including vocabulary size and complexity, total training time, and the available support during testing (e.g., answer options list, inter-phoneme and inter-word intervals controlled by participants). Table II presents an overview of the various works in this area.

Using a discrete-based encoding, Tan et al. [30] demonstrated an average word identification accuracy of 77% on a 251-word set after 5.3 h of training. When the set was expanded to 500 words, the average accuracy dropped to 62% even with additional 1.3 h of training (6.6 h in total), suggesting that recognition is dependent on training of each word. In addition, since 90% of words in their word set were composed of 1–3 phonemes, it is unclear whether participants would be able to generalize their haptic phoneme knowledge to the understanding of untrained and more complex vocabulary.

To the best of our knowledge, this work represents the first assessment of phrases rendered haptically without any kind of additional support. Although some works relying on vocoders [24] [31] have explored the rendering of entire sentences, their devices were intended to function as a complement to lip-reading, for which participants were also presented with video recording of the speaker's face. Approaches based on a discrete mapping have not explored the rendering of entire phrases.

In the final phase of Block 3, in which participants controlled the pace of word rendering, 42% of phrases and 65% of words were correctly identified. Analysis of the extent by which contextual information from identified words in a phrase support overall word identification accuracy was inconclusive. Some participants demonstrated improved accuracy when the words were presented in a phrase (Block 3), perhaps benefiting from the additional 30 min of training, but others performed better on single words in isolation (Block 2), possibly due to the increased difficulty of the word set used in phrases.

Under the more realistic conditions of Block 4, in which phrases were rendered with a fixed IWI, participants achieved a modest average word accuracy of 41%. Since we observed similar accuracy scores between untrained and trained words, the drop in performance can be explained by the participants' lack of experience ($\mu = 21\,\mathrm{min}$) with this testing condition. This required word decoding at a much faster pace and longer retention of the previous elements (i.e., comprehended and partially comprehended words and phonemes) in working memory, resulting in higher cognitive load.

Given the limited training exposure to the final conditions in

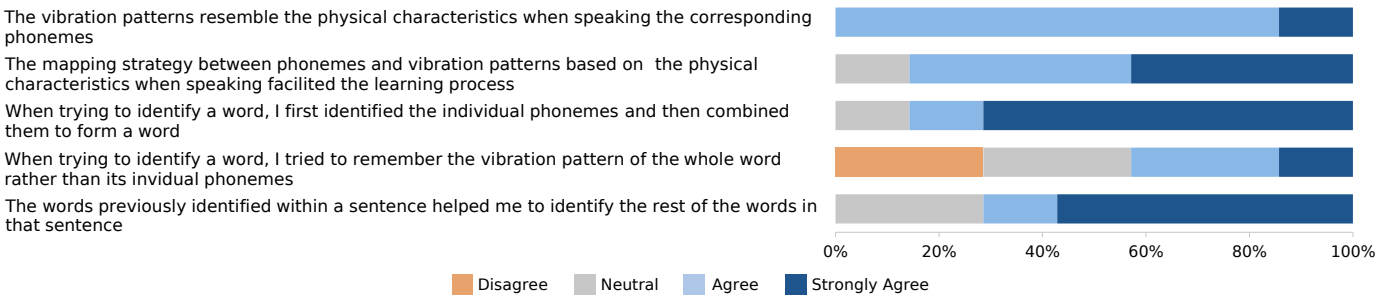

Fig. 10: Post-questionnaire responses—understanding haptic words and phrases.

Block 4, we emphasize that these scores do not represent a final performance metric of the device. Rather, they demonstrate the feasibility of rendering complex phrases through the sense of touch with minimal training time: 4.2 h in total and 2.7 h, ignoring the initial 90 min training (on average) in Study 2 required for participants to regain their performance from the end of Study 1. This training time is several times shorter than works employing vocoders, limited to the rendering of words (55 h on a 150-word list [7]), and comparable with works relying on a discrete mapping (1.7 h on a 100-word list [28], 6.6 h on 500 words [30]).

We expect that after extensive training, stimuli sequences may be decoded as chunks corresponding to longer lexical structures (e.g., syllables, morphemes), similarly to natural language acquisition [32]. This would allow shorter IPI and IWI, leading to decreased cognitive load and facilitating overall comprehension, as mentioned informally by participants.

The observed large individual difference across participants also indicates the need for a more flexible and user-centered training protocol, rather than the fixed structure applied in this study. A compelling strategy is to impose certain minimum scores on all post-tests before allowing participants to advance to the next training stages, rather than using a time-limited approach. This would guarantee as much training time as needed under a manageable level of difficulty before advancing to more complex tasks. Spreading training over a longer period would also leverage the memory consolidation effect.

We hypothesise that with such a training protocol, participants could achieve sufficiently high accuracy on the identification of entire phrases composed of a complex vocabulary, allowing them to participate in conversations receiving speech solely as vibrations delivered to the forearm.

V. Study 3 – Speech-to-Haptic Conversation

Haptic speech replacement shows promise in mediating conversation in situations where traditional sensory modalities to convey linguistic content may be infeasible, such as in the cockpit of an airplane where sensory information may saturate the visual and auditory capacities of a pilot. One of the most popular modes of computer mediated conversation is through messaging apps. Motivated by the results of our previous studies, we thus considered a haptic messaging app as an ideal starting point for a working prototype of a system to mediate haptic conversation between interlocutors. We evaluate the user experience of haptic conversation after minimal training—including how interlocutors adjust their language when speaking haptically, circumstances when haptic communication is most useful, and practical design considerations for such an app. In contrast with many other in-lab studies, this study elicited spontaneous speech and semi-structured bidirectional conversation between participants, and thus provides insight to how such a system may ultimately be used for communication in more natural circumstances.

A. Participants

Three top performing participants from the previous two studies (P2, P4, P8, hereafter "haptic listeners"), as well as two naive participants with knowledge of speech science were recruited two months after the completion of Study 2. All six unique pairings of haptic listeners and naive participants performed the study. Having each naive participant run sessions with all three haptic listeners allowed us to understand how the former learnt the system over repeated sessions.

B. Protocol

The study was conducted over two consecutive days. On the first day, the haptic listeners were allowed 30 min to review the haptic phonemes and practice phrase identification, performing Blocks 3 and 4 activities from Study 2, as per their preference. The second day, each pair of naive participant and haptic listener were placed in separate rooms to conduct the study. Haptic listeners wore headphones playing masking pink noise, with the apparatus attached to their arm as in the previous studies. The naive participants spoke into the messaging app, which translated their utterances to patterned vibrations. These were rendered on the arms of the haptic listeners, who in turn, replied by text message to the naive speaker.

The participants were asked to cooperate to complete several communication tasks together, such as figuring out a time to see a movie or ordering food for a restaurant. These tasks were chosen to overcome a gap in information or collaboratively reason about a problem, using open, unrestricted language.

We recorded the number of turns to complete a task, text of exchanged messages, response time, UI telemetry (e.g., the number of times a certain button was clicked, timing between clicks), and qualitative notes of user experience. To assess communication accuracy, we also asked the haptic listeners to confirm what they "heard" by typing as text the contents of

TABLE II: Comparison of haptic speech rendering approaches from the literature

Apparatus			Training			Testing		
Author (year)	# Actuators (location)	Basic Unit (# symbols)	Time (h)	Set Size	Set Details	Support	Accuracy Untrained	Accuracy Trained
Brooks & Frost (1983) [7]	16 (forearm)	SPC	55	150 W	-	None	-	80% W[a]
Brooks et al. (1985) [23]	16 (forearm)	SPC	80	250 W	-	None	-	76% W[a]
Eberhardt et al. (1990) [24]	1 (table)	SPC	17	40 PHR	6–9 SYL/PHR	Lipreading	33% W	-
Galvin et al. (1999) [25]	8 (fingers)	SPC	12	31 W	-	None	-	70–80% W[a]
Novich (2015) [26]	27 (back)	SPC	11 days[b]	50 W	1 SYL/W	CS (4)	30% W	35–65% W
Liao et al. (2016) [9]	4 (wrist)	L (26)	1	26 L	-	None	-	86% L
Luzhnica et al. (2016) [10]	6 (hand)	L (26)	5	98 W	2–5 L/W	Replay	-	94% L
Zhao et al. (2018) [27]	6 (forearm)	PHO (9)	0.5	20 W	2–3 PHO/W	None	55% W	76% W
Reed et al. (2018) [12]	24 (forearm)	PHO (39)	1.9	39 PHO	-	CS (39)	-	86% PHO
Jiao et al. (2018) [28]	24 (forearm)	PHO (39)	1.7	100 W	2–3 PHO/W	CS (100)	-	81% W
Turcott et al. (2018) [29]	12 (forearm)	PHO (10)	0.8	20 W	2–3 PHO/W	CS (10)	-	76% W
Dunkelberger et a [13]	4 (arm)	PHO (23)	1.7	150 W	1–6 PHO/W ($\mu = 3.1$)	CS (12) USR-IPI	-	87% W
Luzhnica & Veas (2019) [11]	7 (hand)	L (26)	4.7	98 W	2–5 L/W	Replay	-	96% L[c]
de Vargas et al. (2019) [14]	2 (forearm)	PHO (24)	1.7	150 W	1–6 PHO/W ($\mu = 3.1$)	CS (12) USR-IPI	-	94% W
					2–6 PHO/W ($\mu = 3.1$)	None	39% W	51% W
Tan et al. (2020) [30]	24 (forearm)	PHO (39)	6.6	500 W	1–5 PHO/W ($\mu = 2.9$)	None	-	62% W
Block 3 of this work	2 (forearm)	PHO (24)	2.2–3.7	150 PHR (292 W)	2–6 W/PHR ($\mu = 3.6$) 31% SUB	USR-IWI	65% W	65% W
Block 4 of this work			2.7–4.2	225 PHR (514 W)	4–8 W/PHR ($\mu = 5.6$) 42% SUB	None	30–59% W	-

CS: closed set SPC: spectral-based L: letter W: word PHO: phoneme PHR: phrase SUB: words with substitute phoneme(s)
USR-IPI: user-controlled inter-phoneme interval USR-IWI: user-controlled inter-word interval [a] threshold [b] 300 trials/day [c] Lev. dist.

what they thought they received from their conversation partner (CP). Each session concluded with an informal interview with all participants to assess their experience.

C. Results and discussion

We found that interlocutors had to adjust their conversation style in order to effectively communicate using the system. During early sessions, naive participants would speak using long phrases, using many phatic expressions, closely resembling how they would speak during live in-person conversations (e.g. "How are you?...Great, do you want to play a board game?"). Over time, their language became more direct, with less words, often dropping the subject of their clause (e.g. "Friday night menu?"). The mean number of words in phrases exchanged in the first session was 5.47, with a max message length of 9 words. By the last session, participants were exchanging messages with a mean of 3.08 words. The haptic listeners' phonological accuracy—their ability to understand the utterances based off of their precise phonemic constituents—was measured by the phonological edit distance between what the naive speaker said and the haptic listeners' confirmations. Here, we modify this measure from that used in Study 1 so that a value of 1.0 is a perfect match, and 0 is a mismatch. This was achieved by modifying the normalizing expression as the distance between the target string and empty string, and subtracting the result from 1, as follows:

$$1 - \min\left(\frac{pEditDistance(str1, str2)}{pEditDistance(str1, \text{""})}, 1\right) \quad (1)$$

Across all sessions, haptic listeners achieved a phonological accuracy of 0.73. There was an upward trend over time, with the mean accuracy increasing from 0.49 in the first session to 0.92 in the final session. Due to our small sample size, we are unable to confirm whether this is due to the individual capabilities of the haptic listeners or the communication methods of the naive speakers becoming more effective over time. Despite limitations of phonological accuracy, participants were still able to determine the gist of what their CP was saying, as 87.5% of all communication tasks were successfully completed. Many participants noted that the system was best

suited for more information-centered communication, as it was imperative to be concise and difficult to convey suprasegmental features, such as tone of voice, in their messages. S1 remarked, "[this is best when] you already have a real conversation happening in some other context... and this is like a confirmation check", further emphasizing "[you wouldn't use it to] ask, 'what do you want to do?"' S2 noted differing strategies of abbreviation: "[in a conventional messaging app] I would normally say 'Mon' for Monday, but then it's weird to say Mon... Instead of using abbreviations or shortenings I just took out full words". Participants noted some difficulty in the lack of such aspects as tone of voice in a phoneme-based haptic encoding. This loses both emotional and semantic aspects of their speech, which, for example, masks the distinction between an upward tone at the end of the phrase indicating a question, and a static or downward tone indicating a statement.

Haptic listeners tended to replay messages frequently with the "play" button, with an average 3.36 replays during the first session, reducing to 1.88 replays in the last session. H3, a haptic listener, said "longer sentences were harder for me to get in one go... by the time you get to the end of the sentence you forget". Supplying novice haptic listeners with functionality to replay, or memory aids to help recollect earlier elements of the phrase, may thus be an important design factor.

Regarding the encoding system, S1 commented "It's a good setup from a theoretical linguistics perspective... but linguistics isn't necessarily based off how people learn about stuff" noting that the average English speaker does not possess significant phonetic awareness, giving the system a steeper learning curve. H3, a native Hindi speaker, noted that Hindi is typically taught in terms of its articulatory phonetics, and felt that this background offered him an advantage.

VI. Conclusion and Future Work

We have presented the findings from a series of three studies where participants learned how to use WhatsHap, and ultimately completed a simple communicative task that elicited spontaneous speech and semi-structured bidirectional conversation using the device. The training process involved learning a set of 24 vibration patterns representing English phonemes, recognizing words from a sequence of haptic phonemes, and getting habituated to the rendering of entire phrases. After only 4.2 h of training, participants were able to generalize their phoneme identification skills to the understanding of untrained vocabulary, reaching an average word accuracy score of 65% on phrases presented with a user-controlled inter-word interval, and even more encouraging, up to 59% when phrases were rendered with an inter-word interval of 3 s. Three top-performing participants also engaged in communicative tasks in which they had to overcome a gap in information or collaboratively reasons about a problem using open, unrestricted language, receiving a conversation partner's speech entirely as vibrotactile actuation. Despite their limitations on phonological accuracy, participants were able to determine the gist of what their conversation partner was saying, as 87.5% of all communication tasks were successfully completed.

The analysis of communication on these semi-structured bidirectional conversations using WhatsHap also revealed limitations in the ability of the encoding system to represent aspects of emotion or prosody in the message. There is also a question of how to optimize the encoding system for perceptually improved rendering and response times. As one of the participants reported, a morpheme- or syllable-based encoding system may speed up the response times. In languages such as English, having a distinct tacton for each morpheme may lead to an unnecessarily large haptic lexicon. However, using the current phoneme rendering system, we could imagine delivering stimuli rapidly with only large delays between morpheme or word boundaries, and aim to train participants to recognize morphemes instead of phonemes, letting them implicitly infer any phonemic constituents. An alternative encoding could combine a *phonetics* based approach with a *phonology* based approach. Numerous haptic speech rendering approaches, including our own, map a discrete set of vibrotactile symbols to a set of symbolic characters representing the speech sounds. Instead, in a phonetics-based encoding, the patterned vibrations would be based on physical aspects of the speech signal, similar to early efforts on vocoder-based haptic speech replacement, or to more recent systems that translate the audio waveforms directly to vibration through multichannel haptic vests [26]. A benefit of a phonetics based approach is that it preserves suprasegmental aspects of speech associated with emotion and tone. Incorporating acoustic aspects of the speech signal into our encoding system may address some of the limitations revealed—for example, F0 tracking could help disambiguate yes-no questions from statements.

Another promising direction of future research involves training naive users with self-administrated activities personalized to their skill level over an extensive period (e.g., 50 h). The simple design of WhatsHap allows deployment of "haptic speech learning kits" that users could use in the comfort of their homes. This would allow investigation of our encoding strategy in terms of word recognition accuracy and delivery rate, as we seek to achieve communication rates suitable for face-to-face interactions in real-world scenarios. Finally, a study could be run with people with hearing disabilities but preserved phonological awareness to identify features useful for this population and inform the redesign of the apparatus.

References

[1] S. Brewster and L. M. Brown, "Tactons: Structured tactile messages for non-visual information display," in *Proceedings of the Fifth Conference on Australasian User Interface*, vol. 28. Australian Computer Society, Inc., 2004, pp. 15–23.

[2] S. Ertan, C. Lee, A. Willets, H. Tan, and A. Pentland, "A wearable haptic navigation guidance system," in *Second Intl. Symposium on Wearable Computers (Cat. No. 98EX215)*. IEEE, 1998, pp. 164–165.

[3] L. M. Brown and T. Kaaresoja, "Feel who's talking: using tactons for mobile phone alerts," in *CHI'06 extended abstracts on Human factors in computing systems*, 2006, pp. 604–609.

[4] J. R. Cauchard, J. L. Cheng, T. Pietrzak, and J. A. Landay, "Activibe: design and evaluation of vibrations for progress monitoring," in *Proc. Human Factors in Computing Systems*, 2016, pp. 3261–3271.

[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, no. 1, pp. 247–257, 1985.

[6] F. Sorgini, R. Caliò, M. C. Carrozza, and C. M. Oddo, "Haptic-assistive technologies for audition and vision sensory disabilities," *Disability and Rehabilitation: Assistive Technology*, vol. 13, no. 4, pp. 394–421, 2018.

[7] P. Brooks and B. J. Frost, "Evaluation of a tactile vocoder for word recognition," *Journal of the Acoustical Society of America*, vol. 74, no. 1, pp. 34–39, 1983.

[8] S. Engelmann and R. Rosov, "Tactual hearing experiment with deaf and hearing subjects," *Exceptional Children*, vol. 41, no. 4, pp. 243–253, 1975.

[9] Y.-C. Liao, Y.-L. Chen, J.-Y. Lo, R.-H. Liang, L. Chan, and B.-Y. Chen, "Edgevib: Effective alphanumeric character output using a wrist-worn tactile display," in *Proceedings of the 29th Annual Symposium on User Interface Software and Technology*, 2016, pp. 595–601.

[10] G. Luzhnica, E. Veas, and V. Pammer, "Skin reading: Encoding text in a 6-channel haptic display," in *Proc. Intl. Symposium on Wearable Computers*. ACM, 2016, pp. 148–155.

[11] G. Luzhnica and E. Veas, "Optimising encoding for vibrotactile skin reading," in *Proc. Human Factors in Computing Systems (CHI)*. ACM, 2019, pp. 1–14.

[12] C. M. Reed, H. Z. Tan, Z. D. Perez, E. C. Wilson, F. M. Severgnini, J. Jung, J. S. Martinez, Y. Jiao, A. Israr, F. Lau, K. Klumb, R. Turcott, and F. Abnousi, "A Phonemic-Based Tactile Display for Speech Communication," *IEEE Trans. Haptics*, pp. 1–1, 2018.

[13] N. Dunkelberger, J. Sullivan, J. Bradley, N. P. Walling, I. Manickam, G. Dasarathy, A. Israr *et al.*, "Conveying language through haptics: a multi-sensory approach," in *Proc. Intl. Symposium on Wearable Computers*. ACM, 2018, pp. 25–32.

[14] M. F. de Vargas, A. Weill-Duflos, and J. R. Cooperstock, "Haptic speech communication using stimuli evocative of phoneme production," in *2019 IEEE World Haptics Conference (WHC)*. IEEE, 2019, pp. 610–615.

[15] H.-Y. Yao and V. Hayward, "Design and analysis of a recoil-type vibrotactile transducer," *J. of the Acoustical Society of America*, vol. 128, no. 2, pp. 619–627, 2010.

[16] M. A. Mines, B. F. Hanson, and J. E. Shoup, "Frequency of occurrence of phonemes in conversational english," *Language and speech*, vol. 21, no. 3, pp. 221–241, 1978.

[17] J. E. Flege, "Second language speech learning: Theory, findings, and problems," *Speech perception and linguistic experience: Issues in cross-language research*, vol. 92, pp. 233–277, 1995.

[18] A. Chrabaszcz and K. Gor, "Quantifying contextual effects in second language processing of phonolexically ambiguous and unambiguous words," *Applied Psycholinguistics*, vol. 38, no. 4, p. 909–942, 2017.

[19] W. D. Marslen-Wilson and A. Welsh, "Processing interactions and lexical access during word recognition in continuous speech," *Cognitive psychology*, vol. 10, no. 1, pp. 29–63, 1978.

[20] J. Nerbonne and W. Heeringa, "Measuring dialect distance phonetically," in *Computational Phonology: Third Meeting of the ACL Special Interest Group in Computational Phonology*, 1997, pp. 11–18.

[21] K. C. Hall, B. Allen, M. Fry, S. Mackie, and M. McAuliffe, "Phonological corpustools, version 1.2.[computer program]," *Available from PCT GitHub page*, 2016.

[22] R. Vallat, "Pingouin: statistics in python," *The Journal of Open Source Software*, vol. 3, no. 31, p. 1026, Nov. 2018.

[23] P. Brooks, B. Frost, J. Mason, and K. Chung, "Acquisition of a 250-word vocabulary through a tactile vocoder," *The Journal of the Acoustical Society of America*, vol. 77, no. 4, pp. 1576–1579, 1985.

[24] S. P. Eberhardt, L. E. Bernstein, M. E. Demorest, and M. H. Goldstein Jr, "Speechreading sentences with single-channel vibrotactile presentation of voice fundamental frequency," *The Journal of the Acoustical Society of America*, vol. 88, no. 3, pp. 1274–1285, 1990.

[25] K. L. Galvin, P. J. Blamey, M. Oerlemans, R. S. Cowan, and G. M. Clark, "Acquisition of a tactile-alone vocabulary by normally hearing users of the tickle talker™," *The Journal of the Acoustical Society of America*, vol. 106, no. 2, pp. 1084–1089, 1999.

[26] S. D. Novich, "Sound-to-touch sensory substitution and beyond," Ph.D. dissertation, 2015.

[27] S. Zhao, A. Israr, F. Lau, and F. Abnousi, "Coding tactile symbols for phonemic communication," in *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. ACM, 2018, pp. 1–13.

[28] Y. Jiao, F. M. Severgnini, J. S. Martinez, J. Jung, H. Z. Tan, C. M. Reed, E. C. Wilson, F. Lau, A. Israr, and R. Turcott, "A Comparative Study of Phoneme-and Word-Based Learning of English Words Presented to the Skin," in *Human Haptic Sensing and Touch Enabled Computer Applications*. Springer, 2018, pp. 623–635.

[29] R. Turcott, J. Chen, P. Castillo, B. Knott, W. Setiawan, F. Briggs, K. Klumb, F. Abnousi, P. Chakka, F. Lau, and A. Israr, "Efficient evaluation of coding strategies for transcutaneous language communication," in *Intl. Conf. on Human Haptic Sensing and Touch Enabled Computer Applications*. Springer, 2018, pp. 600–611.

[30] H. Z. Tan, C. M. Reed, Y. Jiao, Z. D. Perez, E. C. Wilson, J. Jung, J. S. Martinez, and F. M. Severgnini, "Acquisition of 500 english words through a tactile phonemic sleeve (taps)," *IEEE Trans. Haptics*, 2020.

[31] L. E. Bernstein, M. E. Demorest, D. C. Coulter, and M. P. O'Connell, "Lipreading sentences with vibrotactile vocoders: Performance of normal-hearing and hearing-impaired subjects," *The Journal of the Acoustical Society of America*, vol. 90, no. 6, pp. 2971–2984, 1991.

[32] D. Hewlett and P. Cohen, "Word segmentation as general chunking," in *Proc. Computational Natural Language Learning*. Association for Computational Linguistics, 2011, pp. 39–47.

Maurício Fontana de Vargas is a PhD candidate at the McGill School of Information Studies, currently working on novel Assistive and Augmentative Communication (AAC) systems. His research interests span the field of human computer interaction, including assistive technologies, intelligent user interfaces, and smart environments. Before joining McGill, he received a M.Sc. degree in Electrical Engineering from Federal University of Rio Grande do Sul (UFRGS), and a B.E. in Computer Engineering from State University of Ponta Grossa (UEPG), Brazil.

David Marino is a MSc. student in the department of Electrical and Computer Engineering at McGill University. They are interested in designing natural language interfaces between social agents: human, robot, pet, or otherwise. Research interests include multimodal communication, and designing for emergent and subsymbolic aspects of language. Prior to their studies at McGill, David completed a BA in Cognitive Systems from the University of British Columbia. They are a student member of the Center for Interdisciplinary Research in Music Media and Technology (CIRMMT), and a UBC Language Sciences affiliate member.

Antoine Weill--Duflos received the M.Sc. and Ph.D. in Advanced Systems and Robotics from University Pierre and Marie Curie, Sorbonne Universities, Paris, France, in the Institut des Systèmes Intelligents et de Robotique. He also received the Mechanical Engineering degree (M.Eng.) from the top-ranked French Engineer School Arts et Métiers. He is currently pursuing a postdoctoral fellowship with the McGill University's Shared Reality Lab. His current research interests include haptic device design, robotics, haptic perception and virtual reality. He is a student member of the Centre for Interdisciplinary Research in Music Media and Technology (CIRMMT), and IEEE.

Jeremy R. Cooperstock Jeremy Cooperstock (B.A.Sc. 1990, M.Eng. 1992, Ph.D. 1996), is a professor in the Department of Electrical and Computer Engineering at McGill University. He is an associate editor of the Journal of the Audio Engineering Society, recipient of the award for Most Innovative Use of New Technology from ACM/IEEE Supercomputing and a Distinction Award from the Audio Engineering Society, best paper awards from Transactions on Haptics, Mobile and Ubiquitous Systems, and Haptics Symposium. His research interests focus on multimodal immersive systems and distributed computer-mediated communication. He is an IEEE member and founding member of the Centre for Interdisciplinary Research in Music, Media and Technology.