WhatsHap: A wearable phoneme encoder for real-time haptic conversation

David Marino[1,3][0000−0003−4541−3936],
Maurício Fontana de Vargas[2,3][0000−0002−1594−8318],
Antoine Weill–Duflos[1,3][0000−0002−2782−3678], and
Jeremy R. Cooperstock[1,3][0000−0002−3392−2541]

[1] Department of Electrical and Computer Engineering, McGill University, Montreal, Canada {dmarino,antoinew,jer}@cim.mcgill.ca
[2] School of Information Studies, McGill University, Montreal, Canada
mauricio.fontanadevargas@mail.mcgill.ca
[3] Centre for Interdisciplinary Research in Music Media and Technology, Montreal, Canada

Abstract. WhatsHap is a vibrotactile instant messaging system capable of supporting real-time conversation through the delivery of haptic representations of English phonemes to the arm. The system demonstrates promise in supporting information-centered communication in contexts where there is shared knowledge between users, or where visual or auditory communication is infeasible due to reasons such as sensory saturation. We present an overview of WhatsHap, and an initial analysis of how users learn to communicate haptically with the system.

Keywords: Haptic Communication · Phonology · Multimodal Communication

1 Introduction

Natural language is inherently multimodal, although most frequently received through audition, as in spoken speech, and vision, as in text messaging or sign language. There are, however, many circumstances where both audition and vision may be infeasible. For example, reading a text message while walking near a busy intersection can prove dangerous if visual and auditory senses are preoccupied. Similarly, airline pilots may reach visual and auditory sensory saturation in the cockpit. It is thus useful to be able to encode natural speech in alternative modalities, such as touch, to offload such sensory saturation.

We present *WhatsHap*, a real-time, wearable, alternative communication device capable of speech replacement through dual-channel vibrotactile representations of discrete English phonemes. The vibrotactile symbol system is motivated by the place and manner of articulation of exemplary phones, and is capable of rendering haptic representations of 24 of the 38-40 major phonemes of North American English. The device requires a mobile phone with a headphone jack, connecting to two vibrotactile actuators that are placed on the forearm. A user can speak into their phone's microphone, or send a text message, and the constituent phonemes of their message can then be encoded as haptic symbols that are sent to a listener. Conversely, a user can "listen" to a message from their conversation partner via patterned vibrations on the skin. Other popular linguistic encoding systems, such as Morse code, represent phrases based on their orthography. Delivering a message based instead on its phonology, offers a couple advantages: for languages such as English with a deep orthography, one grapheme can map to many phonemes For example, a $\langle c\rangle$ may be an /s/ as in $\langle \mathbf{c}icada\rangle$ or a /k/ as in $\langle \mathbf{c}at\rangle$. Additionally, multiple graphemes may be required to represent a single phoneme. For example, in the word $\langle morale\rangle$, the -ale maps to /æl/ with the $\langle e\rangle$ unpronounced. This makes an English orthography-based delivery system slower and less transparent.

2 Background

Research on natural tactile communication has demonstrated the feasibility of using the skin as a communication channel for speech. The best-known example is the Tadoma method, in which deaf-blind users receive speech by placing a hand on the talker's face and identifying the physical manifestations during speech production such as lip movements, oral airflow, and throat vibration. Highly trained users were able to achieve identification accuracy scores of 55% for consonants and vowels and 80% for keywords in conversational sentences [12].

In the first attempts to create non-invasive prosthetic devices that would, analogously to the Tadoma method, convey speech through haptics (i.e., tactile speech aids), researchers in the 80s and 90s explored the use of vocoders (see Sorgine et al. [15] for an overview). With this approach, the live acoustic speech signal is processed in real time using a bank of band-pass filters to deliver temporal, intensity, or spectral information on a tactile display consisting of a vibrotactile transducer for each filter band. To achieve desirable identification accuracy scores using such an apparatus, participants needed extensive training: hearing subjects trained for 55 hours to achieve 80% accuracy on a set with 150 words [2], while deaf participants trained for 235 hours (spread across 47 weeks) to achieve 90% on a 135-word list [5]. In addition, experiments with commercially available vocoders demonstrated that they could not be used for understanding speech without the support of lipreading [18].

More recent research on this topic tends to focus on creating a discrete set of tactons or symbols that correspond to letters of the alphabet [9,8] or English phonemes [13,3,16,17], and rendering words as a sequence of these discrete symbols. Using MISSIVE [3], a multi-sensory (radial squeeze, lateral skin stretch, vibration) device worn on the upper arm, participants achieved 87% word identification accuracy—with a self-paced rate of phoneme rendering and choosing their responses from a list of words—after 100 minutes of training on a set with 150 words. Similarly, de Vargas et al. [17] reported an accuracy of 94% under the same training and testing protocol and 45% with an open-answer format and a fixed 1 s inter-phoneme intervals within a word, using a two-channel, lightweight apparatus that delivers vibrations resembling the physical characteristics when uttering the phonemes during normal speech.

In these works, the input vocabulary was pre-programmed into the system, and therefore, fixed. In contrast, we are interested in the possibility of obtaining input from a speech recognition module at the front end of the device responsible for converting the speech audio into a text string, and ultimately, to a haptic encoding that can be rendered as a series of discrete tactile stimuli. To the best of our knowledge, no such apparatus has yet been investigated in the literature.

3 System Design

3.1 Overview

The system consists of a two-channel vibrotactile device connected to a smartphone running a messaging app. The vibrotactile device comprises two voice-coil based actuators (Haptuator Mark-I, Tactile Labs, Montreal, Model no. TL002-14-A) [19], secured with armbands on the dorsal side of the user's forearm. One actuator is placed close to their wrist, and the other close to their elbow. The messaging app runs in a web browser (Figure 1), and allows the user to send a haptic encoding of their speech or text message. In the case of speech, the app uses Google Speech-to-Text API[4] to convert the user's utterances into text. It then uses the CMU Pronouncing Dictionary[5] to obtain a phonemic representation of the utterance's words in North American English. The user can then broadcast their utterance as a stream of haptic symbols to all other interlocutors that are also using the system. If using text input, the algorithm follows the

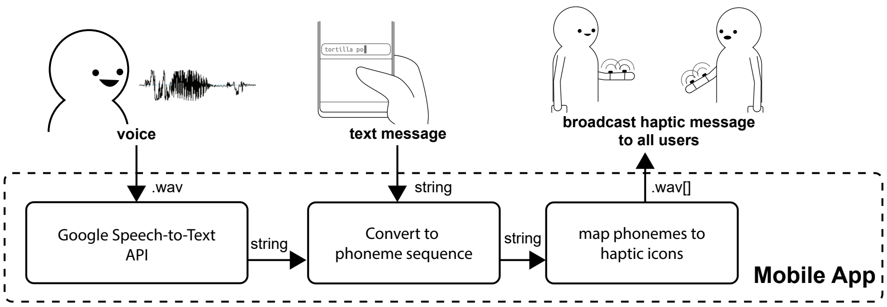


Fig. 1: A sketch of the Mobile App algorithm. The app can take vocal or textual user input, translates their message to a sequence of phonemes using the CMU Pronunciation Dictionary, then outputs the message as a sequence of tactons to all users on the network.

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

same steps excluding speech-to-text conversion. Haptic-encoded speech is then delivered with 1 s inter-phoneme intervals and 3 s inter-word intervals.

3.2 Phoneme-to-Haptic Mapping

Our phoneme-to-haptic mapping follows an encoding system described in a previous study by de Vargas et al. [17]. The haptic stimuli are designed to provide a natural mapping between the haptic sensation and the phoneme's anatomical physicality. The principles are summarized as follows:

Consonants We created haptic symbols representative of 15 English consonants (/p, b, t, d, k, g, f, v, ð, s, z, m, n, l, ɹ/) by augmenting raw audio recordings of exemplary phones.[6] These augmentations were meant to enhance distinctions in *place* and *manner* of articulation. For instance, to express distinctions in place of articulation —e.g., /b/ is pronounced at front of the mouth with the lips, while /g/ is pronounced at the back of the oral cavity— front and back consonants are mapped to distal and proximal regions of the forearm respectively.

Vowels Five vowels, /i, ɛ, ʌ, u, ɑ/, including four dipthongs /eɪ, ɑɪ, ɑʊ, oʊ/ are represented. Audio of the vowels were synthesized in Praat [1]. The vowels were chosen to be synthesized for better control over length and F0. Vowels are given a linear fade effect at phoneme boundaries. As with the consonants, the vowels were mapped to regions of the arm. Dipthongs were represented by a fade between two vowel boundaries.

4 Evaluation and Methods

We aim to understand the design space of our system in terms of the experience and process of conversing using a haptic phonology. Such areas of investigation include *how* users input linguistic content to the system, the user *experience of understanding* a message encoded in haptic phonemes, limitations of our system, and design considerations. We recruited three participants who performed well in a previous study that used a similar haptic encoding system, achieving word identification accuracy scores of 91%, 85%, and 68%, respectively, and two naive participants with experience in speech science. The expert users played the role of a "haptic listener", receiving messages from a conversation partner (CP) in the form of haptic phonemes, while the naive users (or "speakers") communicate using only speech or text. During the main phase of the study, both users were asked to converse using the system through the framework of a joint communication task.

[6] obtained from www.jbdowse.com/ipa

4.1 Description of communication tasks

We created two communication activities based on the concept of task-based language learning [4], an approach extensively used in second language instruction. These tasks focus on achieving a certain goal, rather than a pure linguistic outcome, and have a gap (information, reasoning, or opinion[11]) to be overcome by the conversation partners (CPs).

In the first task (information-gap), the haptic listener played the role of a chef and received a timetable containing a cooking plan for the week, including missing ingredients for each recipe. The speaker's tasks were to find what was on the menu and what ingredients, including quantities, were needed for specific days of the week.

In the second task (reasoning-gap), the speaker was requested to invite the CP to perform an activity, e.g., play golf, and to schedule it at a time both were available according to the timetables they received.

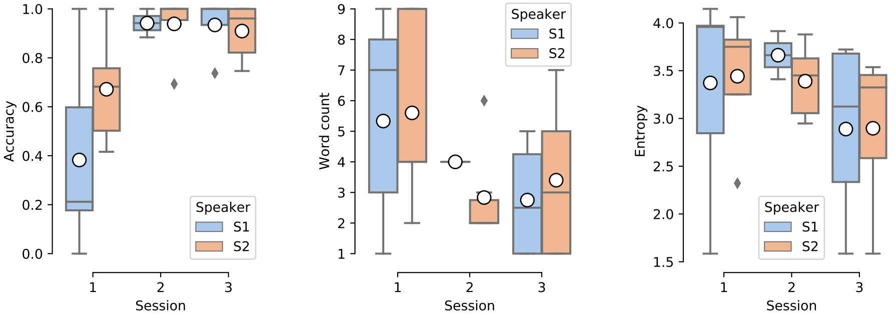
Each of the experts completed two sessions, each involving a separate communication task with a different speaker. The speakers returned for successive sessions with different listeners for a total of three sessions. By structuring our study as such, we gain an understanding of how the naive speakers learn to communicate effectively over repeated sessions, and some appreciation for individual variations between the trained expert listeners. The latter were given a 50 min review session to refresh themselves on the haptic encoding methodology prior to the study proper.

4.2 Analysis and Insights

Our low sample size lacks the statistical power needed to make inferences about the user population as a whole. It does nonetheless provide a description of how individual users performed, and thus, offers insight as to how the design could be improved.

Response time between CPs tended to be longer than natural communication, with each turn lasting an average of 2.84 min ($SD_{ResponseTime} = 1.90\,\mathrm{min}$). Messages were replayed by listeners an average of 2.45 times ($SD_{replay} = 2.01$). Listener accuracy was calculated as a function of how much phonological content was successfully understood, according to the phonological edit distance [6] between the original message and its reiterated interpretation by the listener, defined as the Levenshtein edit distance, i.e., the number of operations required to transform one string to another, weighted by the physiological features [7] associated with each phoneme. We normalized the edit distance as a value between 0 and 1 expressed as $1 - \min(pEditDistance(str1, str2)/pEditDistance(str1, \text{“”}), 1)$, and thus, $1 =$ a perfect match, $0 =$ complete mismatch. Haptic listeners exhibited average accuracy scores of 0.73, with a clear upward trend in accuracy over time (Fig. 2a). A greater sample size is needed to determine if this trend is attributable to the individual skills of the haptic listeners, or because the non-haptic speakers learnt how to adjust their language to communicate effectively using WhatsHap. Regardless of individual phoneme accuracy, participants as a whole were able to comprehend the essence of what their CP was saying, as 87.5 % of all communication tasks were successfully completed. In the first session, to fully comprehend the speaker's intentions, haptic listeners replayed messages a mean of 3.36 times, with the max number of repetitions for a single message being 8 times. By the final session, listeners only replayed messages 1.88 times on average, with the maximum amount of replays for a single message being 3.

We operationalize complexity through the information entropy, i.e., the number of bits needed to represent a string [14], of the messages sent by the naive users, S1 and S2. As they learnt more about the system, the average information entropy, along with the number of words per message, decreased over time, as seen in figures 2c and 2b. There was also a reduction in variance in both of these variables. Decreased complexity exhibits an inverse relationship with accuracy scores.



(a) Accuracy (b) Word count (c) Information entropy

Fig. 2: Boxplots of the evolution of messages sent by naive speakers over the three sessions. Accuracy scores in 2a are based on the normalized edit distance for the reiterated phrase compared to that of the naive user (speaker). White dots indicate the mean.

During the first session, S1 tended to use long sentences ($\bar{X}_{entropy} = 3.37$ bits, $\bar{X}_{words} = 5.33$, $SD_{words} = 3.19$). Their first strategy to improve comprehension was to speak slower and clearer into the microphone, with minimal changes to the words or structure of their sentences. This strategy was sub-optimal because the system delivers haptic phonemes to users at constant intervals regardless of the suprasegmental characteristics of the input speech. During successive sessions, both naive non-haptic participants learned to use much simpler language. S2 mentioned that they tended to stick to pure content words such as nouns and verbs, dropping function words, e.g., articles and prepositions, to convey their points succinctly. In the first session S2's utterances would resemble natural speech with phatic expressions "How are you?... Great, do you want to play a board game?", while only one session later, the utterances became more direct, with pleasantries removed, e.g. "What ingredients?" In later sessions, both participants tended to drop the subject of their clauses, e.g. "want to play golf Sunday?", "Friday night menu?"). By the final session, S1's utterances had reduced to an average of 2.75 words per message ($SD_{words} = 1.78, \bar{X}_{entropy} = 2.88$ bits).

Participants felt that WhatsHap in its present state seemed best suited for information-centered, non-emotional communication. S1 noted "[WhatsHap would be most useful if] you already have a real conversation happening in some other context where you're preferably in a better mode of communication with the other person and this is like a confirmation check", further highlighting that information-focused activities such as conveying grocery lists would be best suited for this system over casual emotional conversations.

5 Discussion and Conclusion

WhatsHap is a messaging app capable of facilitating remote information-centered haptic conversation. The system delivers linguistic content as a sequence of haptic symbols with fixed intervals. This results in long transmission times, for which the time spent to formulate replies is indicative of the high cognitive load (CL) of the activity. However, it bears emphasis that our haptic listeners had only undergone four hours of training, spread over two blocks, performed 14 and 2 months prior to the present study, which is orders of magnitude less than other linguistic tasks involving high CL for novices, such as reading. Thus, we would anticipate both CL and transmission time to decrease with further training.

The discrete, time-constant nature of the encoding system obfuscates distinctions in stress that have semantic consequences, e.g., "*I* didn't notice anything yesterday" vs. "I didn't notice anything *yesterday*. Pitch is also lost, which among its many functions can facilitate distinction between a question and order, e.g., "Let's do 7?" vs. "let's do 7". The use of WhatsHap with text input similarly suffers loss of information since text-based gestures such as emoji are not rendered. Vocal prosody in text is often communicated via methods such all-caps for shouting (HEY YOU), omitting punctuation and alternating capitalization (ReAlLy FuNnY) for sarcasm [10]. These are patterns that the texting community developed over time as they adapted to the medium. If haptic communication was used consistently for everyday conversation, we may expect to see the emergence of similar methods of conveying prosodic cues through touch as speakers gain an understanding of how to play within the confines of the system.

Future directions for WhatsHap can explore how to best facilitate learning haptic communication through effective representation of speech. H3, a haptic-listener who speaks Hindi said that this system may suit some languages more than others because of cultural practice around language instruction and orthography. H3 mentioned that Hindi orthography is taught in terms of its articulatory phonetics, with glyphs having a regularized, close correspondence to speech sounds, unlike English. As a result, he felt that his cultural background made learning the system easier. H3 suggested that presenting a fast sequence of haptic symbols grouped by morphemes may be an ideal method of conveying linguistic content without relying purely on phonology. Timing between speech sounds is a major factor in their perception, so a system that closely resembles the phonetics of an utterance may also prove a useful iteration.

References

1. Boersma, P., Weenink, D.: Praat: Doing phonetics by computer (version 6.0. 37). Retrieved 14.03. 2018 from http://www.praat.org/ (2018)
2. Brooks, P., Frost, B.J.: Evaluation of a tactile vocoder for word recognition. Journal of the Acoustical Society of America **74**(1), 34–39 (1983)
3. Dunkelberger, N., Sullivan, J., Bradley, J., Walling, N.P., Manickam, I., Dasarathy, G., Israr, A., et al.: Conveying language through haptics: a multi-sensory approach. In: Proc. Intl. Symposium on Wearable Computers. pp. 25–32. ACM (2018)
4. Ellis, R., et al.: Task-based language learning and teaching. Oxford University Press (2003)
5. Engelmann, S., Rosov, R.: Tactual hearing experiment with deaf and hearing subjects. Exceptional Children **41**(4), 243–253 (1975)
6. Hall, K.C., Allen, B., Fry, M., Mackie, S., McAuliffe, M.: Phonological corpustools, version 1.2.[computer program]. Available from PCT GitHub page (2016)
7. Hayes, B.: Introductory phonology, vol. 32. John Wiley & Sons (2011)
8. Luzhnica, G., Veas, E.: Optimising encoding for vibrotactile skin reading. In: Proc. Human Factors in Computing Systems (CHI). pp. 1–14. ACM (2019)
9. Luzhnica, G., Veas, E., Pammer, V.: Skin reading: Encoding text in a 6-channel haptic display. In: Proc. Intl. Symposium on Wearable Computers. pp. 148–155. ACM (2016)
10. McCulloch, G.: Because Internet: Understanding the new rules of language. Riverhead Books (2019)
11. Prabhu, N.S.: Second language pedagogy, vol. 20. Oxford University Press (1987)
12. Reed, C.M., Rabinowitz, W.M., Durlach, N.I., Braida, L.D., Conway-Fithian, S., Schultz, M.C.: Research on the Tadoma method of speech communication. The Journal of the Acoustical society of America **77**(1), 247–257 (1985)
13. Reed, C.M., Tan, H.Z., Perez, Z.D., Wilson, E.C., Severgnini, F.M., Jung, J., Martinez, J.S., Jiao, Y., Israr, A., Lau, F., et al.: A phonemic-based tactile display for speech communication. IEEE Transactions on Haptics **12**(1), 2–17 (2018)
14. Shannon, C.E.: A mathematical theory of communication. Bell system technical journal **27**(3), 379–423 (1948)
15. Sorgini, F., Caliò, R., Carrozza, M.C., Oddo, C.M.: Haptic-assistive technologies for audition and vision sensory disabilities. Disability and Rehabilitation: Assistive Technology **13**(4), 394–421 (2018)
16. Turcott, R., Chen, J., Castillo, P., Knott, B., Setiawan, W., Briggs, F., Klumb, K., Abnousi, F., Chakka, P., Lau, F., Israr, A.: Efficient evaluation of coding strategies for transcutaneous language communication. In: Intl. Conf. on Human Haptic Sensing and Touch Enabled Computer Applications. pp. 600–611. Springer (2018)
17. de Vargas, M.F., Weill-Duflos, A., Cooperstock, J.R.: Haptic speech communication using stimuli evocative of phoneme production. In: 2019 IEEE World Haptics Conference (WHC). pp. 610–615. IEEE (2019)
18. Weisenberger, J.M., Percy, M.E.: The transmission of phoneme-level information by multichannel tactile speech perception aids. Ear and hearing **16**(4), 392–406 (1995)
19. Yao, H.Y., Hayward, V.: Design and analysis of a recoil-type vibrotactile transducer. The Journal of the Acoustical Society of America **128**(2), 619–627 (2010)