Towards a Smart Wearable Tool to Enable People with SSPI to Communicate by Sentence Fragments

Gyula Vörös[1], Anita Verő[1], Balázs Pintér[1], Brigitta Miksztai-Réthey[1], Takumi Toyama[2], András Lőrincz[1](✉), and Daniel Sonntag[2]

[1] Eötvös Loránd University, Faculty of Informatics, Pázmány Péter Sétány 1/C,
Budapest 1117, Hungary
{vorosgy,mrb,lorincz}@inf.elte.hu, anitaveroe@gmail.com, bli@elte.hu

[2] German Research Center for Artificial Intelligence, Trippstadter Strasse 122,
67663 Kaiserslautern, Germany
{takumi.toyama,sonntag}@dfki.de

Abstract. The ability to communicate with others is of paramount importance for mental well-being. In this paper, we describe an interaction system to reduce communication barriers for people with severe speech and physical impairments (SSPI) such as cerebral palsy. The system consists of two main components: (i) the head-mounted human-computer interaction (HCI) part consisting of smart glasses with gaze trackers and text-to-speech functionality (which implement a communication board and the selection tool), and (ii) a natural language processing pipeline in the backend in order to generate complete sentences from the symbols on the board. We developed the components to provide a smooth interaction between the user and the system thereby including gaze tracking, symbol selection, symbol recognition, and sentence generation. Our results suggest that such systems can dramatically increase communication efficiency of people with SSPI.

Keywords: Augmentative and alternative communication · Smart glasses · Eye tracking · Head-mounted display · Speech synthesis · Natural language processing · Language models

1 Introduction and Related Work

The ability to communicate with others is one of the most basic human needs. People with severe speech and physical impairments (SSPI) face enormous challenges during seemingly trivial tasks, such as shopping. A person who cannot

Author to whom correspondence should be addressed. This work was carried out as part of the EITKIC 12-1-2012-0001 project, which is supported by the Hungarian Government, managed by the National Development Agency, financed by the Research and Technology Innovation Fund and was performed in cooperation with the EIT ICT Labs Budapest Associate Partner Group.

P. Cipresso et al. (Eds.): MindCare 2014, LNICST 100, pp. 90–99, 2014.
DOI: 10.1007/978-3-319-11564-1_10

speak may be able to communicate directly to his closest relatives only, relying on them completely to interact with the world [1].

Understanding people using traditional alternative and augmentative communication (AAC) methods – such as gestures and communication boards – requires training [2]. These methods restrict possible communication partners to those who are already familiar with AAC.

In AAC, utterances consisting of multiple symbols are often telegraphic: they are unlike natural sentences, often missing words to speed up communication [3].

Some systems allow users to produce whole utterances or sentences that consist of multiple words. The main task of the AAC system is to store and retrieve such utterances [4]. However, using a predefined set of sentences restrict the things the user can say severely. Other approaches allow generation of utterances from an unordered, incomplete set of words [5–7], but they use predefined rules that constrain communication.

The most effective way for people to communicate would be spontaneous novel utterance generation – the ability to say anything, without a strictly predefined set of possible utterances [8].

We attempt to give people with SSPI the ability to say almost anything. For this reason, we would like to build a general system that produces novel utterances without predefined rules. We chose a data-driven approach in the form of statistical language modeling.

In some conditions (e.g., cerebral palsy), people suffer from communication and very severe movement disorders at the same time. For them, special peripherals are necessary. Eye tracking provides a promising alternative to people who cannot use their hands [9].

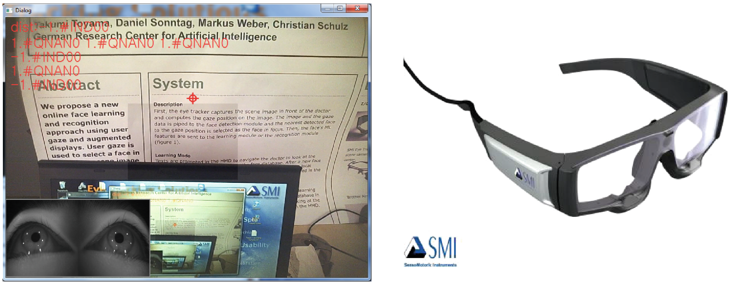
We also attempt to enable communication for people with SSPI *almost anywhere*. We identify two subproblems blocking this aim: (i) to overcome the barriers experienced by people with SSPI and (ii) to help other, non-trained people understand them easily. In our proposed solution to the first problem, smart glasses with gaze trackers (thereby extending [15]) and text to speech (TTS) play the role of a communication board, a selection tool, and the bridge to the environment. For the second subproblem, we propose a statistical language model based approach to generate complete sentences from the selected symbols. Finally, utterances can be synthesized by a text to speech system.

We developed the components of smooth interaction between the smart glasses and the user, including gaze tracking, symbol selection, symbol recognition, and sentence generation. In light of recent technological developments, we expect that the complete system will fit on a pair of smart glasses – making whole sentence communication possible anywhere – in the very near future. Furthermore, it has been shown that adaptive optimization is feasible in the similar case of head motion controlled mouse cursor [10] and that ability-based optimization can have considerable advantages [11].

2 Components and Tests

We used the following tools:

- Eye Tracking Glasses by SensoMotoric Instruments GmbH (in the following, ETG): a pair of glasses with eye tracking infrared cameras and a forward looking video camera (Fig. 1).
- AiRScouter Head Mounted Display by Brother Industries (in the following, HMD): a see-through display which can be attached to glasses.
- MPU-9150 Motion Processing Unit by InvenSense Inc. (in the following, MPU): a device with integrated motion sensors.


Dialog
Takumi Toyama, Daniel Sonntag, Markus Weber, Christian Schulz
German Research Center for Artificial Intelligence
Abstract
System
SMI
SMI

(a)Eye tracking software.

(b)Eye tracking glasses.

Fig. 1. System used to track eye movements. The glasses contain a forward looking camera, and two eye cameras. The latter capture images of the eyes (shown on the bottom left corner of the screenshot), which are illuminated by six infrared light sources.

The system has three main functions: gaze tracking, symbol selection and utterance generation (Fig. 2). Our components implement different aspects of this system. We performed four tests, one for each component.

In gaze tracking, a crucial problem is calibration. The user can adapt to the errors of the system up to a point, but as the error grows, the system is more and more difficult to use, and calibration is needed. In the first test, (Sect. 2.1) we consider this calibration problem during symbol selection. We use the ETG and the HMD to perform gaze tracking and display the symbols. This test also serves as a demonstration of symbol selection with an easily portable system, as the communication board is on the HMD.

In the tests of the second and third components, we simulated a higher resolution HMD – necessary to display the communication board – with a projector and a board made of paper. In the second test, (Sect. 2.2) the participant is communicating with his communication partner in artificial scenarios (e.g., shopping). This component combines the ETG and a projector. An MPU is used to initiate gaze calibration.

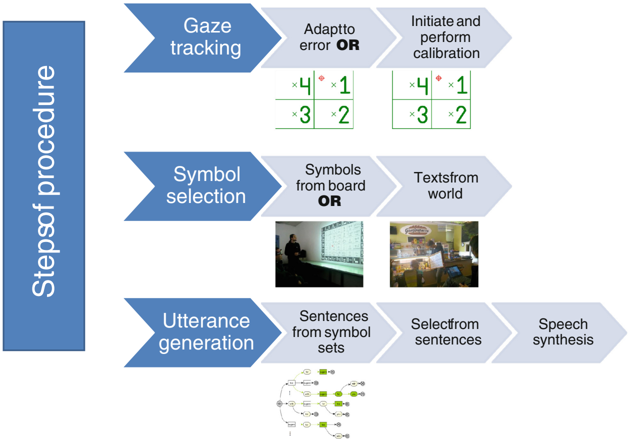


Fig. 2. Main components of the full system, and the tests.

The third test takes communication out into the real world (Sect. 2.3). Participants are shopping in a real shop. Optical character recognition is used to recognize texts which can be added to the communication table as symbols.

The final component is natural language sentence generation (Sect. 2.4). The selected symbols are assembled into natural language sentences using statistical language modeling. We used the transcripts produced in the second test to test this component. In the proposed system, sentence generation will be interactive: the user will be able to select the sentence to be uttered by TTS from a list of the most probable candidate sentences.

2.1 Symbol Selection and Calibration on Head Mounted Display

We found the size of the HMD too small for symbol-based communication. We performed a series of tests to study the effects of gaze tracking error in a situation with proper symbol size. In these tests, a simple "communication board" of 4 symbols was used, as this was within the limitations of the technology.

Participants wore ETG with an HMD. A small red crosshair showed the estimated gaze position of the participant. The area of the HMD was split into four rectangular sections. The sections were numbered from 1 to 4 (see the top of Fig. 2). The goal was to select the numbers in increasing order by keeping the red crosshair on the appropriate rectangle for two seconds. Each selection was confirmed with an audible signal. After selecting number 4, the numbers were set in a new arrangement. The objective of the participants was to make as many correct selections as they could in a fixed amount of time.

To study errors in the calibration, the position of the red crosshair was translated by a small, fixed amount of artificial error in every five seconds, so the overall error increased by time. Participants tried to adapt to the erroneous crosshair position by compensating with their gaze. When the error grew too large, the participant could signal to ask for a reset. This removed the artificial error, but it also prevented the users from selecting the next number for five seconds. The timeframe of one test was 80 s, which is long enough to perform a fair number of selections (usually 40–60), and allows for an artificial error so big that we think no one can tolerate.

There were four participants; they did not have SSPI. After getting used to the system, each of them did four tests. There were a total of 18 resets. The average amount of artificial gaze error in the instant of a reset was 120 pixels, which corresponds to approximately 2.7° field of view. The results indicate that the participants are able to tolerate relatively large errors in gaze tracking.

In the following, we describe two scenarios with full-sized communication boards (Sects. 2.2 and 2.3). They are followed by the description of sentence fragment generation (Sect. 2.4).

2.2 Communication with Predefined Symbols

The participant of these tests, B., is a 30 years old man with cerebral palsy. He usually communicates with a headstick and an alphabetical communication board, or with a PC-based virtual keyboard controlled by head tracking.

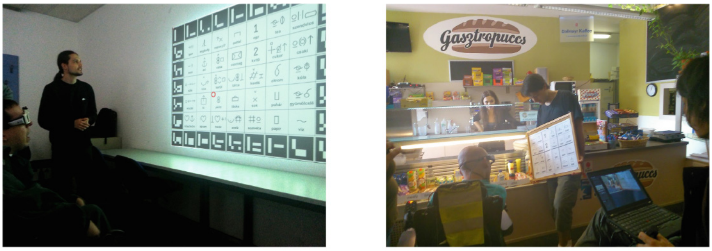
During the tests, a communication board with Bliss symbols and Hungarian labels was projected on a wall (Fig. 3a). The size of the board is reasonable for HMDs coming to the market. The participant sat in front of the projected board, wearing ETG. The gaze position on the board was calculated using fiducial markers. The estimated gaze position was indicated as a small red circle on the projected board (similarly to the previous test). A symbol was selected by keeping the red circle on it for two seconds.

The eye tracking sometimes needs recalibration; the participant could initiate recalibration by raising his head straight up. This was detected by the MPU. Once the recalibration process was triggered, a distinct sound was played, and an arrow indicated where the participant had to look.

The tests had two scenarios. In the first one, the participant wished to buy food in a store; a communication partner played the role of the shop assistant. The communication board was designed for that situation, and contained 35 Bliss symbols. In the second situation, the participant and his partner discussed appointments (e.g., times and places to meet). This involved another communication board with 42 Bliss symbols. In day-to-day situations, communication boards could be easily switched by next/previous symbols.

To verify that communication really happened, the participant indicated misunderstandings using their usual yes-no gestures, which were quick and reliable. Moreover, a certified expert in AAC was present, and indicated apparent communication problems.

We found that the error rate was small: of the 205 symbol selections that occurred, only 23 was mistaken, which means approximately 89 % accuracy. The errors were corrected by the participant in the test. This error rate is typical when our participant uses communication boards.



(a) Predefined symbols.

(b) External symbols.

Fig. 3. Communication with symbols. In the test with predefined symbols (a), the participant is wearing eye tracking glasses. The communication partner is standing. The communication board is projected on the wall. The black symbols around the board are fiducial markers. The white symbols each denote a single Hungarian word. The small red circle provides feedback about the estimated gaze position. The tests with external symbols (b) took place in an actual food store. Here, the communication board is printed on a piece of paper. Optical character recognition is used to let the participant combine the symbols on the board and symbols in the external world.

2.3 Communication with External Symbols

In real-life situations, the appropriate symbols may not be present on the communication board. The user's environment, however, may contain words which can be recognized using optical character recognition (OCR). It would be very useful if the system could recognize these texts, read them out loud, and include them in the utterances the user can construct. To test the possibilities of communication with ETG in a real-life scenario, we performed tests in an actual food store with a board made of paper (Fig. 3b).

The two participants have cerebral palsy. Both of them are unable to speak, and they also cannot read, but they understand spoken language.

The communication board was printed on paper. We had also placed labels near the items in the store, with the name of the item on them. We used OCR and speech synthesis to recognize and read out loud words in the environment and on the communication board, based on gaze position. As the OCR system sometimes did not recognize the text under our light conditions, a person watched a computer screen showing the forward-looking camera image with the participant's gaze position, and read out loud the words (from the board or from the labels) the participant was looking at, to simulate a better OCR.

During the tests, the communication partner understood the intentions of the participants, and they were able to communicate. To verify this, the same methods were used as in Sect. 2.2.

2.4 Sentence Fragment Generation

The natural language processing component of the system generates sentence fragments from the chosen symbols. Each symbol corresponds to an open class word (i.e., nouns, verbs, etc.). The algorithm works by producing permutations of these input words and a predefined set of closed class words (i.e. prepositions, etc.). The probabilities of the permutations are estimated with a language model. In a live system, a small number of the most probable permutations (i.e., sentence fragments) will be presented to the user to choose from.

As the number of permutations grows exponentially, we developed a greedy algorithm that builds the sentence fragments step-by-step by traversing a pruned prefix tree of word sequences (i.e., ngrams) based on their probabilities. It is similar to the Beam search algorithm [12], but it constrains the search space to reduce computation time instead of reducing the memory footprint.

We traverse the prefix tree breadth-first, starting from the first word of the sentence fragment. The root of the tree is a start symbol <S>: the parent of the possible sentence fragments. Each branch in the tree is a possible sentence fragment. A node is expanded by adding input or closed words to it. Pruning is based on two thresholds: one for the minimum acceptable probability of a fragment, and one for the maximum number of children on each expanded node. We always keep the children of a node with the highest probabilities. Figure 4 shows a step of the algorithm for two input words and a tiny closed word set.

To further restrict the search space, we use only those co-occurrent closed words which are most likely to follow each other, avoiding a huge amount of unintelligible combinations like *"the the"*, *"an the"* etc., but using common co-occurrent closed words e.g., *"would have"*, *"next to the"*.

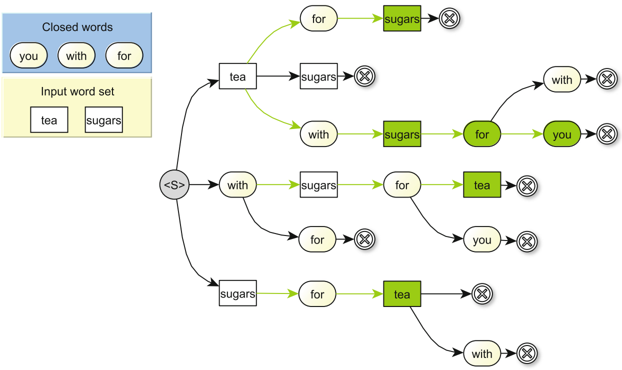


Fig. 4. The algorithm in action, with a tiny closed word set. Rectangle shaped nodes are the elements of the input word set: {*tea, sugars*}. Nodes with rounded shape correspond to closed words. Nodes with green color indicate the first word of an accepted fragment. A branch which contains all the input words can contain more fragments, like the results *"tea with sugars"*, *"tea with sugars for"* and *"tea with sugars for you"* on branch {*tea, with, sugars, for, you*}. The X symbol means that the branch has no expanded leaves, as its estimated probability falls below the threshold. It has been cut and will not be expanded any further.

We used language models generated from two corpora. The Google Books N-gram corpus [13] is very large but also very general: it was compiled from the text of books scanned by Google. The OpenSubtitles2013 corpus [14] is a compilation of film and television subtitles[1], closer to spoken language. We discarded duplicates based on their video identifier and their content. The language of the live tests was Hungarian, but the language models are English. In this simple setting, translation was straightforward.

We examined conversation records of the food shopping scenario (Sect. 2.2) as they contained the most natural language communication. In this first study, we were interested in the feasibility of the method: can a data-driven system work in this complex scenario? The test shows that the answer is yes: there are already a number of cases where the method can help the participant tremendously. Some examples are included in Table 1.

Table 1. Sentence fragments generated from *word sets*. The fragments with the four highest scores are shown in the table, with the highest scoring on the top. The fragment deemed correct is shown in bold.

Symbols	The generated sentence fragments	
	OpenSubtitles	Google Books
tea, two, sugars	two sugars and tea	**tea with two sugars**
	tea with two sugars	tea and two sugars
	tea for two sugars	sugars and two tea
	and two sugars tea	tea for two sugars
tea, lemon, sugar	**tea with lemon and sugar**	**tea with lemon and sugar**
	lemon tea and sugar are	tea with sugar and lemon
	one lemon tea and sugar	lemon and sugar and tea
	sugar and tea with lemon	tea and sugar and lemon
would, like, tea	**I would like some tea**	instead of tea would like
	I would like to tea	tea no one would like
	would you like some tea	everything would you like tea
	I would like the tea	no one would like tea
one, glass, soda	**one glass of soda**	soda and one glass
	no one glass soda	glass of soda one
	no one soda glass	soda in one glass
	and one soda glass	soda to one glass

3 Discussion

We proposed a system to enable people with SSPI to communicate with natural language sentences. We demonstrated the feasibility of our approach in four tests

[1] www.opensubtitles.org

of components of the system. We plan more tests with the full system, in real situations, when the available technology makes it possible.

In light of recent technological developments, we expect that the complete system can be realized in the very near future. In fact, the whole system could fit on a pair of smart glasses with:

1. a 3D gaze tracker that estimates the part of 3D space observed by the user
2. a camera monitoring the environment that reads the signs in that volume by natural sign recognition or optical character recognition methods. The signs can be converted to symbols on the communication board
3. a head mounted display to be used as the communication board, a gaze calibration tool and a feedback tool about the precision of the gaze tracker
4. natural language processing to transform series of symbols to whole sentences within the context
5. TTS to transform the sentences to utterances

It may be possible to handle a communication board larger than the HMD: the user could look at different parts of the board using motion sensors to track his head movements. We assume that the spoken answer can be understood by the SSPI person. This is not a necessity: answers of the partner can be transformed into symbol series by means of automated speech recognition tools.

4 Outlook

Another direction of improvement besides the integration of improved and miniaturized hardware technology is to improve the algorithms.

We identified three aspects of the system where substantial improvements can be made, namely concerning the described calibration, adaptation and prediction algorithms (Fig. 5). Currently, gaze calibration can be initiated by the user, and the user has to adapt to calibration errors him or herself (until a recalibration step is initiated). A more advanced system could integrate calibration and adaptation to (i) continuously adapt to the user to reduce gaze interpretation errors and (ii) detect when calibration is needed and recalibrate automatically. Similarly for prediction, a smoother interaction may be possible by adapting to

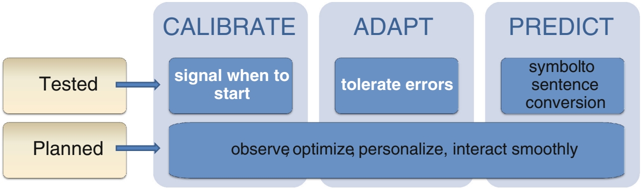

CALIBRATE
ADAPT
PREDICT
Tested
Planned
signal when to start
tolerate errors
symbolto sentence conversion
observe, optimize, personalize, interact smoothly

Fig. 5. Tested and planned solutions to the problems of calibration, adaptation and prediction in the system. Activities with white font are performed by the user; those in black are performed by the system.

the context and the participants of the communication scenario. As people with SSPI have diverse needs, personalization could help tremendously in all three aspects.

References

1. Blackstone, S.W., Dowden, P., Berg, M.H., Soto, G., Kingsbury, E., Wrenn, M., Liborin, N.: Augmented communicators and their communication partners: a paradigm for successful outcomes. In: CSUN Technology and Persons With Disabilities Conference 2001 (2001)
2. Pennington, L., Goldbart, J., Marshall, J.: Interaction training for conversational partners of children with cerebral palsy: a systematic review. Int. J. Lang. Comm. Dis. **39**(2), 151–170 (2004)
3. Wiegand, K., Patel, R.: Non-syntactic word prediction for AAC. In: SPLAT 2012, pp. 28–36 (2012)
4. Arnott, J.L., Alm, N.: Towards the improvement of augmentative and alternative communication through the modelling of conversation. Comput. Speech Lang. **27**(6), 1194–1211 (2013)
5. McCoy, K.F., Pennington, C.A., Badman, A.L.: Compansion: from research prototype to practical integration. Nat. Lang. Eng. **4**(1), 73–95 (1998)
6. Karberis, G., Kouroupetroglou, G.: Transforming spontaneous telegraphic language to well-formed greek sentences for alternative and augmentative communication. In: Vlahavas, I.P., Spyropoulos, C.D. (eds.) SETN 2002. LNCS (LNAI), vol. 2308, pp. 155–166. Springer, Heidelberg (2002)
7. Patel, R., Pilato, S., Roy, D.: Beyond linear syntax: an image-oriented communication aid. Assistive Technol. Outcomes Benefits **1**(1), 57–66 (2004)
8. ASHA: Augmentative and Alternative Communication Decisions. http://www.asha.org/public/speech/disorders/CommunicationDecisions/
9. Calvo, A., Chiò, A., Castellina, E., Corno, F., Farinetti, L., Ghiglione, P., Pasian, V., Vignola, A.: Eye tracking impact on quality-of-life of ALS patients. In: Miesenberger, K., Klaus, J., Zagler, W.L., Karshmer, A.I. (eds.) ICCHP 2008. LNCS, vol. 5105, pp. 70–77. Springer, Heidelberg (2008)
10. Lőrincz, A., Takács, D.: AGI architecture measures human parameters and optimizes human performance. In: Schmidhuber, J., Thórisson, K.R., Looks, M. (eds.) AGI 2011. LNCS, vol. 6830, pp. 321–326. Springer, Heidelberg (2011)
11. Gajos, K.Z., Wobbrock, J.O., Weld, D.S.: Improving the performance of motor-impaired users with automatically-generated, ability-based interfaces. In: CHI 2008, pp. 1257–1266 (2008)
12. Zhang, W.: State-Space Search: Algorithms, Complexity, Extensions, and Applications. Springer, New York (1999)
13. Goldberg, Y., Orwant, J.: A dataset of syntactic-ngrams over time from a very large corpus of English Books. In: SEM 2013, vol. 1, pp. 241–247
14. Tiedemann, J.: Parallel data tools and interfaces in OPUS. In: LREC 2012, pp. 23–25 (2012)
15. Sonntag, D., Zillner, S., Schulz, C., Weber, M., Toyama, T.: Towards medical cyber-physical systems: multimodal augmented reality for doctors and knowledge discovery about patients. In: Marcus, A. (ed.) DUXU 2013, Part III. LNCS, vol. 8014, pp. 401–410. Springer, Heidelberg (2013)