Open Journal of Modern Linguistics, 2015, 5, 291-301
Published Online June 2015 in SciRes. http://www.scirp.org/journal/ojml
http://dx.doi.org/10.4236/ojml.2015.53026

The Microsoft KINECT: A Novel Tool for Psycholinguistic Research

Rinus G. Verdonschot[1], Héloïse Guillemaud[2], Hobitiana Rabenarivo[2], Katsuo Tamaoka[3]

[1]Waseda Institute for Advanced Study, Waseda University, Tokyo, Japan
[2]Graduate School of Engineering, Nagoya University, Nagoya, Japan
[3]Graduate School of Languages and Cultures, Nagoya University, Nagoya, Japan
Email: rinusverdonschot@gmail.com

Received 29 May 2015; accepted 26 June 2015; published 30 June 2015

Abstract

The Microsoft KINECT is a 3D sensing device originally developed for the XBOX. The Microsoft KINECT opens up many exciting new opportunities for conducting experimental research on human behavior. We investigated some of these possibilities within the field of psycholinguistics (specifically: language production) by creating software, using C#, allowing for the KINECT to be used in a typical psycholinguistic experimental setting. The results of a naming experiment using this software confirmed that the KINECT was able to measure the effects of a robust psycholinguistic variable (word frequency) on naming latencies. However, although the current version of the software is able to measure psycholinguistic variables of interest, we also discuss several points where the software can still stand to be improved. The main aim of this paper is to make the software freely available for assessment and use by the psycholinguistic community and to illustrate the KINECT as a potentially valuable tool for investigating human behavior, especially in the field of psycholinguistics.

Keywords

Language Production, Psycholinguistics, KINECT, Psychological Research Tool

1. Introduction

The way we interact with technology is rapidly changing. While we were once limited to keyboards and point-and-click devices, we can now interact with technology using our whole body. The rapidly decreasing cost of 3D sensing technologies (such as the Microsoft KINECT), even allows us to interact with technology through facial expressions and voice information. Although this technology offers exciting new opportunities for expe-

How to cite this paper: Verdonschot, R.G., Guillemaud, H., Rabenarivo, H., & Tamaoka, K. (2015). The Microsoft KINECT: A Novel Tool for Psycholinguistic Research. *Open Journal of Modern Linguistics, 5,* 291-301.
http://dx.doi.org/10.4236/ojml.2015.53026

rimental research on human behavior, the actual implementation of these novel technologies is still in its infancy. This paper highlights a potentially important role for KINECT technology in a particular area concerning the study of human behavior, namely language production (a subfield of psycholinguistics).

This paper is structured as follows: First, we provide a brief background on the existing research and theoretical models of language production, and summarize how dependent variables (such as naming latencies and accuracy) are usually obtained. Second, we introduce several important features of the KINECT sensor and review their potential applications within experimental psycholinguistic research. Subsequently, we discuss the C# software developed by our lab (all code freely downloadable), which implements the KINECT device to an experimental paradigm by depicting a characteristic experimental situation found in psycholinguistics. Next, we present experimental data within a genuine experimental setting by testing 34 participants on a word-frequency paradigm by using the KINECT and validate this data by using an established method in the field (i.e., by voice key). Finally, we point out particular shortcomings of the current version of the software and avenues for resolving these shortcomings and implementing the KINECT in future research, both on language production and in general.

1.1. Short Background on Language Production Research

Although the KINECT offers advancements for behavioral research in many fields, this paper focuses on how the KINECT can benefit research on language production (a part of experimental psycholinguistics). Within the language production literature, there are several theoretical models that describe the way speech is produced: starting from ideas in our head and ending with the actual pronunciation of words (e.g. Dell, 1986; Levelt, Roelofs, & Meyer, 1999). Most of the experimental data supporting these models comes from chronometric research (i.e. measuring reaction time latencies) using basic "triggering devices" such as buttons and voice keys (i.e. electronic circuits initiating a pulse if an input volume crosses a certain threshold). Typical experimental paradigms used in language production research either show a particular stimulus on the screen or present a stimulus auditorily and wait for the participant to name a particular target out loud. The time it takes from seeing (or hearing) the stimulus to naming it out loud is called the reaction time (RT) and serves as the main dependent variable together with the accuracy of the response. However, classic lab equipment such as voice keys only capture RTs for the onset of a single word at a time, and the difference between speech and other (irrelevant) sounds (e.g. coughing) cannot be distinguished without time consuming post-hoc (or online) manual response checking (although there is freely available software which substantially eases and optimizes this task such as Check Vocal; Protopapas, 2007). This is because voice key triggering will simply occur if the input volume crosses a certain threshold. Additionally, data will be usually lost if the voice input does not exceed that threshold (e.g. when a participant speaks softly for instance). Moreover, voice keys have no semantic capabilities, which again instigate a need for manual response checking. Finally, some questions have arisen about the reliability of voice keys. For example, when speaking, even after phonemes are produced it may take the voice key varying amounts of time to detect them, since some sounds take more or less time to initiate (e.g. /z/ versus /p/; see Kessler, Treiman, & Mullennix, 2002; Sakuma, Fushimi, & Tatsumi, 1997). It is therefore reasonable to state that paradigms found in experimental psycholinguistics can be limited by particular aspects of experimental equipment.

1.2. The Microsoft KINECT Device

In contrast to devices designed to be implemented for scientific use only, the KINECT is a device (costing roughly 200 USD) developed by Microsoft to be used with video games (e.g. on XBOX and Windows). The KINECT enables users to interact with a computer via gestures and voice commands.

The KINECT (v1)[1] contains an infrared (IR) emitter and IR depth sensor (640 × 480 pixels) for 3D tracking, a RGB camera (1280 × 960 pixels) to acquire high-quality RGB color video (both the IR depth sensor and the RGB camera operate at 30 fps) and a microphone array, which contains four microphones for capturing sound[2] (see **Figure 1**). The IR emitter emits infrared light in a pre-determined "speckle pattern" (which are in fact small dots of infrared light that fall on everything in front of the KINECT camera). The IR depth sensor perceives these patterns and determines depth by looking at the displacement of specific dot patterns (e.g. on objects close to the KINECT the dot pattern will be spread out, but on far objects the dot pattern will be much denser). Addi-

[1]Since summer 2014 there is a new version of the KINECT sensor (v2) with improved specifications.

[2]More information can be found at: http://msdn.microsoft.com/en-us/library/jj131033.aspx.

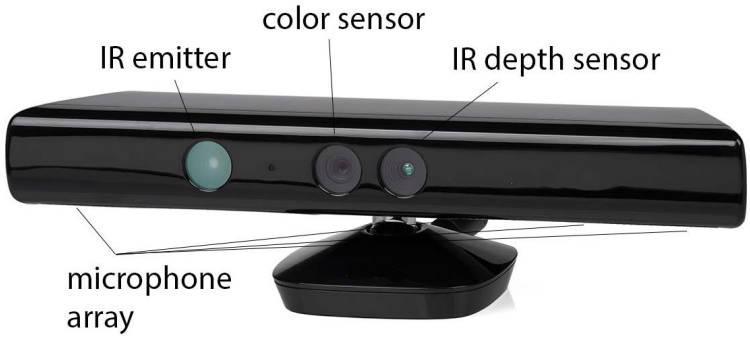

Figure 1. the Microsoft KINECT http://msdn.microsoft.com/en-us/library/jj131033.aspx.

tionally, as there are four microphones, it is possible to accurately retrieve the spatial location of the sound source (e.g. a person speaking), as well as being able to record what is spoken. Furthermore, by using an accelerometer it is possible to determine the current orientation of the KINECT and the integrated tilt motor can be used to track objects or people within the room.

For research in language production, one particularly important feature of KINECT is its ability to track the human face (**Figure 2**). Microsoft has made a so-called Software Development Kit (SDK; current version for KINECT v1 is 1.8) available which contains numerous programming routines to track a human face in real time (downloadable from http://www.microsoft.com/en-us/kinectforwindows/). This SDK can measure roughly 100 points (including so-called "hidden points") resulting in real-time face-tracking. Thus, the KINECT is able to build a detailed model of the human face, called a face mesh, using sets of triangles and lines.

1.3. Opportunities Offered by the KINECT for Research in Psycholinguistics

Naturally, the most important issue for researchers is how the KINECT can contribute to their research. The following list, though incomplete, offers five potential ways we believe the KINECT could advance language production research:

1) The KINECT can track lip movements in real-time, allowing researchers to obtain detailed information on the speech planning process even before actual speech sounds are uttered. By focusing on the distances between particular points on the lips and face, in combination with the speech recognition pack (found in the SDK), it is possible to determine the onset and offset of individual words. In this paper we report our preliminary efforts to build a novel program that detects the detection of the beginning and end of individual words, by tracking lip movements.

2) Another exciting feature is that the KINECT is able to track more than one person over time, which would allow for language experiments to take place in a more natural, conversational setting.

3) The KINECT has the potential to perform basic eye tracking, allowing researchers to assess approximately where participants are looking on a screen. Experimental paradigms may benefit from these additional behavioral measures, which could indicate, for instance, whether participants are engaged in the task at hand, and, if so, which parts of the screen they are mainly fixating on.

4) The KINECT comes with advanced voice recognition (including language packs for many major languages), allowing for automatic post-hoc accuracy checking (see examples in the SDK provided by Microsoft).

5) It has been shown that the KINECT is able to on-line track and interpret body gestures (Biswas & Basu, 2011; Suarez & Murphy, 2012) and basic emotions (Zhang, Zhang, & Hossain, 2013; Piana, Staglianò, Odone, Verri, & Camurri, 2014), allowing for another dimension to be added to the dependent measures in a psycholinguistic experiment.

1.4. A First Attempt to Implement the KINECT into the Area of Language Production

As far as we know, there is no previous language production literature that utilizes the KINECT. This paper therefore represents the first attempt in this field to integrate the KINECT into the daily practice of a psycholinguistics lab. In this paper we focus on implementing the first of the five-abovementioned points, that is, the tracking of lip movements in real-time to gather information on the speech planning process.

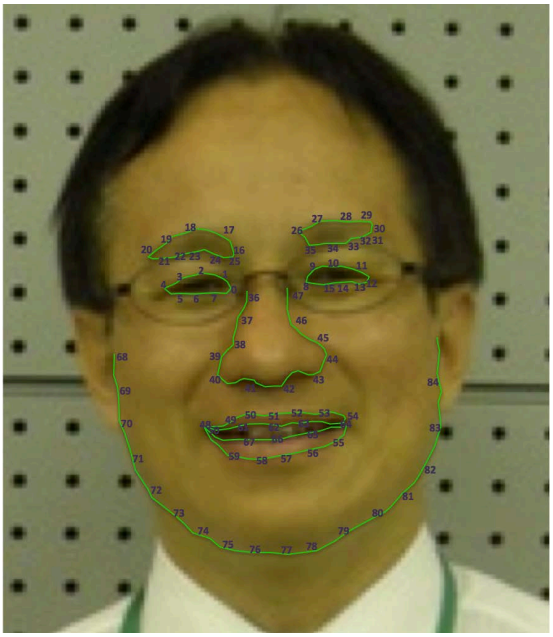

Figure 2. Points on a human face which can be tracked by KINECT[2].

As there are no previous instances for comparison (again, as far as we know), we set out to program a working version of the KINECT software (using C#) to display experimental stimuli and measure a psycholinguistic variable of interest. We aim to keep the code open and freely available for other researchers to use and adapt to their own insights. Obviously, when running the program, an attached KINECT for Windows, including the SDK is required (and Visual Studio is needed when adapting the code). To accommodate those who do not have this setup we provide a short video demonstrating the program online (https://www.youtube.com/watch?v=KCsLuBZeu5c). Furthermore, the program (executable and source code) is provided at: http://www.languageandcognition.nl/rinusverdonschot/kinect.rar. Notice that we provide the complete working directory in this file to have everything available to experienced programmers (for those who simply want to run the program the executable can be found in /bin/x86/debug/FaceTrackingBasics-WPF.exe). The KINECT SDK v1.8 needs to be installed as well.

Although the KINECT is able to track more than one person, in this initial stage of program development, only a single person is tracked during an experiment. The current version of the program is able to:

1) Randomly display a word (taken from an Excel file) to a participant.

2) Use the KINECT to determine the visual on- and offset of the word relative to its initial presentation (i.e. lip/face points) in real-time.

3) Use the KINECT to detect the auditory on- and offset of the word relative to its initial presentation (i.e. sound threshold) in real-time.

4) Generate an output file (Excel), which contains these latencies separated by onset and offset for both visual and auditory modalities.

Concerning the input and output files (Points 1 and 4) it is fair to say that these issues are not complicated and in fact are independent from the KINECT itself. The software we designed uses an Excel sheet as its input. In this file, one column needs to denote the to-be-presented stimuli and another column needs to denote the conditions for the experiment. Stimuli are randomized before presentation (all parameters, such as font type, font size, presentation durations, ISI, etc., can be changed in the C# code). The output file is also an Excel file generated by the software. This output file contains the onset events and latencies both for the visual and auditory modalities. As we provide the source code with no restrictions, we encourage experienced programmers to adjust it to their own liking and introduce additional options for the community. We hope that community efforts will enable the practical implementation of the KINECT into psycholinguistic labs.

Concerning the visual detection of the word (Point 2), it is first important to realize that visual lip detection (as we implemented it) is not identical to lip-reading (which is much more complex). We focused on whether it

was possible to visually track the start and end of words using KINECT's face- and lip-points, in order to assess naming latencies. To do this, we measured the distance from the upper lips to the lower lips and the distance between the corners of the mouth. We found that it was complicated to rely on absolute (direct) distances, mostly because the points provided by KINECT sensor have a tendency to move (as real-time video is never really static). Therefore we also measured the variation of these distances over brief time intervals. We implemented a calibration method to get a standard "non-speaking" person against which a "speaking" person could be compared. We additionally acquired and set several triggers in real-time (detailed information can be found in the C# code) to limit the effect of visual noise on the sensor as much as possible. We found that natural light coming from behind the KINECT- (that is, a natural light source in front of the subject) provided the most accurate tracking results. Lighting should preferably be uniform with as little shadow on the face as possible. Under these conditions, we found the ideal distance from the KINECT to be around 50 - 80 cm. However, if "ideal" conditions cannot be implemented in experimental settings, one should still aim to create uniform lighting in the room. Additionally, distance between the participant and the KINECT can be increased to between 1 m and 1.5 m if the tracking remains poor for particular room conditions. The software provides a simple face mesh (virtual lips) to ascertain whether the tracking is working well (to open this feature press "Display Video" on the main window). To ensure that tracking is working, observe the lip face mesh in the window called "window Showing". The face mesh should not "shake" substantially when the mouth is closed. To provide important real-time technical information for the experimenter during an experiment we added a button called "Display Info" to the main window. This button opens a new window containing the following information:

Dist 3D (Computed Distance): This value represents the opening of the mouth, while taking into account both the vertical (upper lips to lower lips) and horizontal (corners of the lips) axes for the mouth opening. We only use the 3D points directly provided by the KINECT. These points have the advantage of being fairly independent of the distance from the participant to the sensor. This method provides added reliability compared to a measurement that only relies on the vertical opening of the mouth. Additionally, this method is quite valuable when detecting lip movements in the middle of a word (see also the "Number Keys Distances" in the same window). Though adequate for current purposes, this measure stands to be improved in future work to enhance speech onset detection.

Mouth Opening: The initial vertical opening of the mouth used to assess the beginning of a word (i.e. when the mouth goes from a closed position to an open position).

Mouth opening trigger: a dynamic value (see C# code for specific implementation) used to assess the trigger for the beginning of the word using visual input. It is difficult set a pre-determined level for "Mouth Closed" and "Mouth Opened" as this may trigger false positives due to the fact that occasionally responses to movements not followed by a sound may occur.

Number Keys Distances: This value represents the variation of "Dist 3D". One key corresponds to a small set of closing values for "Dist 3D", which are dynamically stored for up to 1 second. The larger the number is at any one given time, the more the lips are moving.

Trigger Distances: The minimum number of keys of "Number keys distances" until the mouth is considered to be "moving enough". In ideal conditions this should always be "1" when the subject's mouth is motionless, but when the conditions of the experiment are somewhat unfavorable (due to lighting etc.), the tracking points may move excessively and the number of keys may reach values greater than 1. Therefore, we built in a calibration method to determine the number of keys when the mouth is not moving (see downloadable C# code for the specific implementation).

Number Keys Mouth Opening: Same as the previous element but this pertains to "mouth opening" and has a fixed value. Sometimes after a participant finishes speaking, s/he will not go back to a "closed position" as defined after the calibration. Therefore, we consider the participant must have stopped speaking when the mouth is not moving anymore. However, this check adds some minor uncertainty to the software (and stands to be improved).

Head Moving: This value is true when the head is moving "excessively": when head movement hampers the face point detection. This value is used to assess whether the information retrieved by the KINECT sensor is reliable.

Sound Energy: This value can be interpreted as the volume of the sound. It is computed using the Root Mean Square amplitude of the recordings gathered by the KINECT's microphone array.

Sound Speaking: This value is true if the software detects that the participant is speaking the presented stimu-

lus word. Otherwise, this value remains false.

Concerning the auditory detection of the word (Point 2) we compute the volume of the sound produced in the environment. When the sound volume reaches a particular level, we consider that the participant must have been speaking. Likewise, when the sound level drops under a certain threshold, we consider that the participant must have stopped speaking. Since online computations take time, and occasionally may trigger false positives (due to environmental noise), we devised another verification method, which made the detection more accurate but does not yet release all uncertainty within the online detection process. Please see the provided C# code for more specific details regarding these implementations.

2. Experiment: Measuring Word Frequency Effects Using the KINECT

To assess whether our software for the KINECT could be implemented in a real, online experimental situation we have chosen to investigate a well-known effect, which has shown itself to be robust across 50 years of experimental work: the word frequency effect (e.g. Oldfield & Wingfield 1965; Jescheniak & Levelt, 1994; Starreveld, La Heij, & Verdonschot, 2013). The word-frequency effect refers to the observation that word-reading latencies are inversely related to the frequency of occurrence of a word in the language.

2.1. Methods

2.1.1. Participants

Thirty-four native Japanese participants (31 female; *M* = 22.5 years old; 1 left-handed) from Yamaguchi University (Japan) took part in this experiment. All participants gave informed consent before the start of the experiment and received a small reward for participation.

2.1.2. Stimuli

We selected 30 high-frequent (HF) and 30 low-frequent (LF) Japanese Kanji words. Word frequencies were calculated by using a corpus containing 11 years of articles from the all-Japanese newspaper, the Mainichi Newspaper, spanning from 2000 to 2010. The morphological parsing program MeCab 0.991 counted 477,264 morphological units (type frequency) and a total token frequency of 299,695,840 from this newspaper corpus. The word frequencies significantly differed between high (*M* = 82,295, *SD* = 55,242) and low frequency (*M* = 960, *SD* = 157) words, t (58) = 8.06, $p < 0.001$. Their natural logs between high (*M* = 11.13, *SD* = 0.61) and low frequency (*M* = 6.85, *SD* = 0.24) words differed significantly, t (58) = 36.08, $p < 0.001$. The words in the two sets were matched with respect to two characteristics. First, the length in moras was controlled between high (*M* = 3.63, *SD* = 0.56) and low frequency (*M* = 3.60, *SD* = 0.50) words, t (58) = 0.25, *ns*). Second, the visual complexity, measured by stroke numbers, was also matched between high (*M* = 18.17, *SD* = 5.55) and low frequency (*M* = 18.00, *SD* = 3.95) words, t (58) = 0.13, *ns*). Please see **Appendix A** for an overview of the stimuli used.

2.1.3. Procedure

Participants were seated in front of a 120 Hz 17.3" LED laptop screen (Alienware Mx 17; GeForce 680 M 2 GB DDR 5 GPU) in combination with a KINECT v1 sensor. Only after the participant's face could be accurately tracked and after assessing whether mouth open and close movements could be registered (using 12 practice trials), would the experiment begin. The initial tracking and calibrating process differed between participants, from person to person and lasted anywhere between 15 seconds and one minute.

Before the experiment started and after each trial was completed, a rapid calibration routine was performed to determine the mouth position and location. Participants were asked to keep their mouth closed during calibration, and when they were not uttering words. After the presentation of a fixation point (+) for 1000 ms the target word appeared on the screen (1.38˚ × 0.69˚ at 60 cm; MS Mincho) and disappeared when the participants named the word or if no word was uttered after 4 seconds. The experimenter recorded whether participants accurately named the word (further software development using novel KINECT's voice recognition libraries will be able to make this process automatic). There was no break between the 60 targets. The entire procedure lasted about 10 - 15 minutes.

2.2. Results

In total 5 participants were excluded from the analysis due to failure to accurately record their lip movements

(i.e. over 50% recording failures) or calibration. We currently have no explanation as to why the KINECT was unable to process the face data for these particular participants (further testing is required). For the remaining 29 participants we discarded lip detection errors and erroneous utterances (2.8% of the data). Additionally, we discarded data points laying outside the 2.5 SD range of a participant's mean per condition for each dependent variable (DV; i.e. 2.1% for visual onset, 1.3% for visual offset, 2.5% for auditory onset, and 1.9% for auditory offset). Response latencies for each dependent measure were analyzed using linear mixed effect (LME) models (e.g., Bates, Maechler, Dai, 2008; Baayen, 2008), according to the lme 4.0 package (from http://lme4.r-forge.rproject.org) available in R (version 3.1.0, R Development Core Team, 2008). We used the lmerTest package in R, which provides LME with p-values using Satterthwaite's approximation for the degrees of freedom (Kuznetsova, Brockhoff, & Christensen, 2014). In the analyses, frequency (i.e., HF vs. LF) was treated as fixed factor. Random factors were by-subject and by-item intercepts. The model used in the analysis of response latencies was the following: DV ~ Frequency + (1 | Participants) + (1 | Items). See **Table 1** for a summary of the experimental results.

These results show that the software in combination with the KINECT indeed is able to record a well-known psycholinguistic variable (i.e. HF < LF RTs), both by using auditory (speech) information and by using visual (mouth movement) information.

3. Comparing the KINECT Findings with Typical Voice Key Latencies

To verify whether a typical language production experiment would show the same pattern of results when using a voice key, we re-ran the same experiment with a new group of students.

3.1. Methods

3.1.1. Participants

Twenty native Japanese participants (14 female; *M* = 21.6 years old) from Waseda University (Japan) took part in this experiment. All participants gave informed consent before the start of the experiment and received a small reward for participation.

3.1.2. Stimuli

Stimuli are identical to the KINECT experiment described in 2.1.2.

3.1.3. Procedure

Participants were tested individually in a quiet room. The experiment was programmed using the DMDX software package (Forster & Forster, 2003), combined with a custom voice key. Participants were asked to read aloud a target word presented on a CRT monitor as quickly and as accurate as possible. The trial order and stimulus properties were similar to the KINECT experiment except that there were no calibration routines necessary.

3.2. Results

Responses were manually corrected for voice key errors via visual inspection of the speech waveforms using the Check Vocal program (Protopapas, 2007). Outliers outside the 2.5 SD range of a participant's mean per condi-

Table 1. Results of the KINECT and voicekey experiments.

DV	HF (SD)	LF (SD)	Effect	DP	t	*p*
V-On	603 (209)	686 (282)	83	1519	3.98	<0.001
V-Off	1372 (363)	1438 (408)	66	1226	2.15	<0.05
A-On	848 (150)	902 (196)	54	1626	3.59	<0.001
A-Off	1486 (196)	1558 (249)	72	1583	3.37	<0.01
VK-On	500 (71)	549 (133)	49	1147	5.35	<0.001

DV = dependent variable, HF = high frequency, LF = low frequency, SD = standard deviation, DP = data points included in the analysis, V = visual, A = auditory, VK = Voice Key On = onset, Off = offset, t = t-value, *p* = p-value.

tion for each dependent variable were discarded (2.5%), as were naming errors (1.9%). Response latencies were analyzed similarly to the KINECT experiment. See the last row of **Table 1** for a summary of the experimental results using the voice key. Results closely matched those found using the KINECT: demonstrating the classic word frequency effect (i.e., participants were quicker to name HF words, compared to LF words). The effect size was comparable between the voice key (49 ms) and KINECT (54 ms), but slightly smaller than the visual detection (83 ms). The overall mean voice key latencies were much faster than the auditory detection and slightly faster than the visual detection (we will come back to this in the discussion).

4. Discussion and Future Prospects

The program depicted in this paper *is not meant to be a definite product nor to represent any leading standard* in using the KINECT for psycholinguistic research. For instance, as can be seen **Table 1**, the t-value for VK-On is higher than for V-On, indicating that it can better account for the word frequency effects' variance. However, this not startling as this paper describes *the first attempt* to implement the KINECT in a typical psycholinguistic research setting.

Although we effectively managed to measure on- and offset naming latencies and the word frequency effect for both the visual (face mesh) and auditory domains, particular aspects stand to be improved upon. We particularly mention three aspects of our present implementation: 1) Visual detection latency: our method detects a variation of face point distances over time. This has the consequence that the particular threshold used to decide whether genuine speech is present actually takes place at the peak of the variation, that is, when the mouth is opened (to a certain degree). Information concerning the real beginning of opening the mouth is actually found in earlier frames, which need to be assessed after threshold detection. This particular computation currently adds some uncertainty to the result (for the exact implementation see the C# code) as to which is the genuine start of the lips movement (especially when the head is moving at the same time). Therefore, although we were able to accurately detect the frequency effect, overall, the visual detection latencies are roughly 100 - 140 ms longer than actual voice key latencies. Although it is not uncommon for different groups to diverge in their mean latencies when the environment was not identical (KINECT and VK experiments were conducted at different locations and with different software), we believe our software can be optimized. Preferably, independent verification on the particular way the beginning of the lip variation is assessed should be undertaken and – if not deemed accurate - improved upon.

Next, 2) Sound detection latency: we only used the microphone array from the KINECT, and the Application Programming Interface (API) for v1 does not provide any direct way to acquire the volume (as far as we can tell). To consider the data concerning the volume threshold to be used in speech detection, we computed the actual volume using the Root Mean Square amplitude. To do this, we needed to accumulate several instances of sound data information, which incurred some detection uncertainty (this may explain the sizable mean latency differences between the A-On and VK-On). However, the next generation of the KINECT sensor (v2) and its SDK might allow for a more direct assessment, which future endeavors will need to investigate. Nevertheless, it is important to recognize that the current software did *measure the frequency variable correctly* within the auditory modality, and further developments will only improve upon the accuracy.

Lastly, 3) it is important to point out that with commercial products hardware characteristics have typically been independently verified by using photocells, oscilloscopes and time-to-live (TTL) pulses. Similarly for the KINECT hardware, accuracy and resolution of the depth system (used for the visual onset/offset in our software) has been heavily investigated since its first release in 2010 (e.g. Khoshelham & Oude Elberink, 2012). However, in the current case, the RTs (experimental results) given by the KINECT largely depend on the specific implementation of the code (and not of the hardware). Therefore, additional software optimization and development is necessary before any thorough validation of the KINECTs physical timing characteristics and its ability to measure RTs in a controlled psycholinguistic research setting can be made.

5. Conclusion

The KINECT is a novel device from Microsoft that has the potential to become a valuable tool in measuring human behavior (such as producing language). The current paper presents a version of the KINECT software that is suitable to measure psycholinguistic variables of interest in a typical language production experiment. The results show that the KINECT is indeed able to measure frequency differences between words in the visual

and auditory modalities. In future investigations, automatic accuracy checking (via KINECT SDK language packs) should be one of the major features added to a new version of the software. Future collaboration among psycholinguistic labs worldwide is necessary to optimize and validate the current software into an advanced product that is suitable for a large variety of experimental situations. This will take time and effort, but the gains will be great.

Acknowledgements

We would like to thank John Phillips at Yamaguchi University, Japan for recruiting participants to test the KINECT, and Wido La Heij at Leiden University, the Netherlands for reading an earlier version of the manuscript. We would also like to thank Masahiro Yoshihara at Waseda University, Japan for recruiting and testing participants. The present work was supported by a Grant-in-Aid for Challenging Exploratory Research from the Japan Society of the Promotion of Science (JSPS) Grant Number 25580112 (principal researcher: Katsuo Tamaoka).

References

Baayen, R. H. (2008). *Analyzing Linguistic Data: A Practical Introduction to Statistics Using R.* Cambridge: Cambridge University Press. http://dx.doi.org/10.1017/cbo9780511801686

Bates, D., Maechler, M., & Dai, B. (2008). Lme4: Linear Mixed-Effects Models Using S4 Classes [Computer Software Manual]. http://lme4.r-forge.rproject.org/

Biswas, K. K., & Basu, S. K. (2011). Gesture Recognition Using Microsoft Kinect®. *Proceedings of the IEEE International Conference on Automation, Robotics and Applications*, Wellington, 6-8 December 2011, 100-103. http://dx.doi.org/10.1109/ICARA.2011.6144864 http://www.ieee.org/conferences_events/conferences/conferencedetails/index.html?Conf_ID=19125

Dell, G. S. (1986). A Spreading-Activation Theory of Retrieval in Sentence Production. *Psychological Review*, *93*, 283-321. http://dx.doi.org/10.1037/0033-295X.93.3.283

Forster, K. I., & Forster, J. C. (2003). DMDX: A Windows Display Program with Millisecond Accuracy. *Behavior Research Methods, Instruments & Computers*, *35*, 116-124. http://dx.doi.org/10.3758/BF03195503

Jescheniak, J. D., & Levelt, W. J. M. (1994). Word Frequency Effects in Speech Production: Retrieval of Syntactic Information and of Phonological Form. *Journal of Experimental Psychology: Language, Memory, and Cognition*, *20*, 824-843. http://dx.doi.org/10.1037/0278-7393.20.4.824

Kessler, B., Treiman, R., & Mullennix, J. (2002). Phonetic Biases in Voice Key Response Time Measurements. *Journal of Memory and Language*, *47*, 145-171. http://dx.doi.org/10.1006/jmla.2001.2835

Khoshelham, K., & Oude Elberink, S. (2012). Accuracy and Resolution of Kinect Depth Data for Indoor Mapping Applications. *Sensors*, *12*, 1437-1454. http://dx.doi.org/10.3390/s120201437

Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. B. (2014). lmerTest: Tests for Random and Fixed Effects for Linear Mixed Effect Models (Lmer Objects of Lme4 Package). R Package Version 2.0-6. http://CRAN.R-project.org/package=lmerTest

Levelt, W. J. M., Roelofs, A., & Meyer, A. S. (1999). A Theory of Lexical Access in Speech Production. *Behavioral and Brain Sciences*, *22*, 1-75. http://dx.doi.org/10.1017/S0140525X99001776

Oldfield, R. C., & Wingfield, A. (1965). Response Latencies in Naming Objects. *Quarterly Journal of Experimental Psychology*, *17*, 273-281. http://dx.doi.org/10.1080/17470216508416445

Piana, S., Staglianò, A., Odone, F., Verri, A., & Camurri, A. (2014). Real-Time Automatic Emotion Recognition from Body Gestures. *Computing Research Repository* (CoRR). http://arxiv.org/abs/1402.5047

Protopapas, A. (2007). Check Vocal: A Program to Facilitate Checking the Accuracy and Response Time of Vocal Responses from DMDX. *Behavior Research Methods*, *39*, 859-862. http://dx.doi.org/10.3758/BF03192979

R Development Core Team (2008). R: A Language and Environment for Statistical Computing. Vienna: R Foundation for Statistical Computing.

Sakuma, N., Fushimi, T., & Tatsumi, I. (1997). Measurement of Naming Latency of Kana Characters and Words Based on Speech Analysis: Manner of Articulation of a Word-Initial Phoneme Considerably Affects Naming Latency. *Japanese Journal of Neuropsychology*, *13*, 126-136.

Suarez, J., & Murphy, R. (2012). Hand Gesture Recognition with Depth Images: A Review. *Proceedings of the IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*, Paris, 9-13 September 2012, 411-417. http://dx.doi.org/10.1109/roman.2012.6343787 http://www.ro-man2012.org/

Starreveld, P. A., La Heij, W., & Verdonschot, R. G. (2013). Time Course Analysis of the Effects of Distractor Frequency and Categorical Relatedness in Picture Naming: An Evaluation of the Response Exclusion Account. *Language and Cognitive Processes*, *28*, 633-654. http://dx.doi.org/10.1080/01690965.2011.608026

Zhang, Y., Zhang, L., & Hossain, A. (2013). Multimodal Intelligent Affect Detection with Kinect. *Proceedings of the 12th International Conference on Autonomous Agents and Multi-Agent Systems*, St. Paul, 6-10 May 2013, 1461-1462.

Appendix A. Stimuli Used in the Experiments

High-Frequency Words			Low-Frequency Words		
相手	aite	Companion	愛用	aiyoo	favorite
安全	anzen	Safety	暗号	angoo	code
代表	daihyoo	representative	弾力	danryoku	elasticity
男性	dansei	Man	台車	daisya	platform truck
打線	dasen	baseball lineup	脱皮	dappi	ecdysis
団体	dantai	organization	同名	doomei	same name
学校	gakkoo	School	害虫	gaityuu	harmful insect
反対	hantai	opposition	白馬	hakuba	white horse
販売	hanbai	Sales	配列	hairetu	arrangement
海外	kaigai	Foreign	歌劇	kageki	opera
開発	kaihatu	development	格段	kakudan	special
監督	kantoku	supervision	回路	kairo	circuit
改正	kaisei	revision	皇帝	kootei	emperor
会社	kaisya	company	完勝	kansyoo	complete victory
過去	kako	the past	河口	kakoo	mouth of river
確認	kakunin	confirmation	仮想	kasoo	imagination
関係	kankei	relation	感度	kando	sensitivity
環境	kankyoo	environment	快速	kaisoku	high speed
家族	kazoku	Family	火薬	kayaku	gunpowder
結果	kekka	Result	建材	kenzai	building material
交渉	koosyoo	negotiations	国策	kokusaku	national policy
強化	kyooka	strengthen	幸運	kooun	good luck
問題	mondai	question	木製	mokusei	wooden
内容	naiyoo	subject	南方	nanpoo	south
来年	rainen	next year	落差	rakusa	a head
削減	sakugen	Cut	最適	saiteki	optimum
写真	syasin	photograph	作風	sakuhuu	literary style
野球	yakyuu	baseball	役場	yakuba	town hall
予算	yosan	estimate	洋画	yooga	Western painting
財政	zaisei	financial affairs	雑草	zassoo	weed

To improve the chance for successful detection we did not employ initial CV (where V= u) moras as they typically evoked very small lip movements.