Read My Lips: Towards Use of the Microsoft Kinect as a Visual-Only Automatic Speech Recognizer

Peter McKay, Bryan Clement, Sean Haverty, Elijah Newton, and Kevin Butler
Department of Computer Science and Information Science, University of Oregon
Eugene, Oregon
{pem,clement,haverty,ebnewt,butler}@cs.uoregon.edu

ABSTRACT

Consumer devices used in the home are capable of collecting ever more information from users, including audio and video. The Microsoft Kinect is particularly well-designed for tracking user speech and motion. In this paper, we explore the ability of current models of the Kinect to support use as an automatic speech recognizer (ASR). Lip reading is known to be difficult due to the many possible lip motions. Our goals were to quantify lip movement while observing the correlation with recognized words. Our preliminary results show that word recognition through the audio interface and with use of the Microsoft Speech API can provide upwards of 90% accuracy over a corpus of words, and that the visual acuity of the Kinect is such that we can capture a total of 22 data points representing the lip model through the Face Tracking API at a high resolution. Based on these results and that of recent work, we forecast that the Kinect has the ability to act as an ASR and that words can potentially be reconstructed through the observation of lip movement without the presence of sound. Such an ability for household devices to observe and parse communication presents a new set of privacy challenges within the home.

1. INTRODUCTION

As consumer devices have become increasingly powerful, they are equipped with an ever-increasing array of sensors for interacting with the external world. Devices capable of capturing audio and video input from users are nearly ubiquitous in homes throughout the industrialized world. The preponderance of these devices makes them effective at generating data for further analysis and to the benefit of a user, but such information can also be used to compromise a user's privacy. One of the challenges, from a consumer's standpoint, is understanding just how capable these consumer devices are of collecting information and at what granularity.

Microsoft's Kinect device has been a well-received addition to the XBox video game system, and allows for many new modes of interaction between a user and the system, along with recognition abilities that can present finer input to a program. The Kinect allows the ability to perform body tracking and to calculate range of motion, thus moving its applicability from the realm of strictly entertainment into areas such as healthcare. Already the Kinect has been used in an orthopedic setting to evaluate foot [10] and body posture [4], and as a tool for providing exercises to further physical rehabilitation [3]. We have found, however, that the Kinect is a suitable device for collecting even finer-grained information from users. Specifically, we postulate that lip tracking and lip reading are possible. There have been numerous rumors that the forthcoming version of the Kinect will enable lip reading [5]; but in this work, we show that the current Kinect already has the potential for tracking lip movement and lip reading, thus providing the potential for acting as an automatic speech recognizer (ASR).

The Kinect is capable of recognizing English words through the speech recognition API. We used this as a basis for investigating its accuracy, by testing recognition of individual words with a speaker positioned various distances from it, and testing accuracy in the presence of background noise. We performed our investigation by evaluating existing code samples and the contents of the available APIs, in an effort to confirm the hypothesis that the Kinect is a valid platform for the desired environmental analysis, and our preliminary results show that a single speaker seated 1 meter or less away from a Kinect can have the speech of a corpus of predefined words captured and recognized approximately 90% of the time. We also find that our cursory models for lip analysis are able to identify 22 discrete vertices representing a user's mouth, allowing a distinct polygonal shape to be extracted. Based on this data, we present a plan of inquiry for how we can build on these results to show the Kinect's ability to track lip movement and to correlate it with spoken words, to eventually allow for lip reading without sound.

The layout of the remainder of the paper is as follows: Section 2 provides background on the Kinect; Section 3 describes our methodology; Section 4 describes our preliminary evaluation; and Section 5 provides our plans for furthering existing results to demonstrate the Kinect's ability to act as an ASR, and how that may affect user privacy in unanticipated ways.

2. BACKGROUND

The Microsoft Kinect provides a wide range of sensing capabilities in a relatively compact form-factor, with a comparatively reasonable price tag. The Kinect presents access

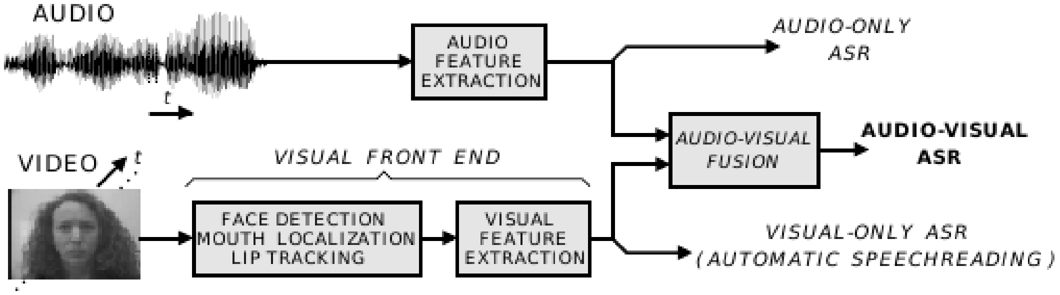
AUDIO
t
AUDIO FEATURE EXTRACTION
AUDIO-ONLY ASR
VIDEO
t
VISUAL FRONT END
AUDIO-VISUAL FUSION
AUDIO-VISUAL ASR
FACE DETECTION MOUTH LOCALIZATION LIP TRACKING
VISUAL FEATURE EXTRACTION
VISUAL-ONLY ASR (AUTOMATIC SPEECHREADING)

Figure 1: The main processing blocks of an audio-visual automatic speech recognizer from [11]. The visual front end design and the audio-visual fusion modules introduce additional challenging tasks to automatic recognition of speech compared to traditional audio-only ASR.

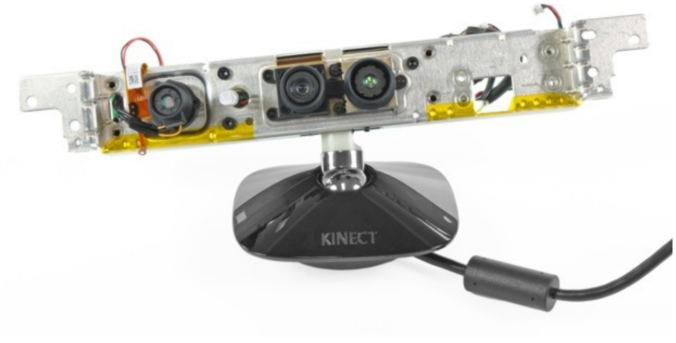

Figure 2: Internals of the Microsoft Kinect [2]

to a four-channel microphone array, as well as two different methods of visual environment monitoring. These visual monitoring devices include a color camera, as well as an infrared rangefinding camera capable of constructing detailed depth maps. In our search of available online documentation, we found much research in the field of audio/visual speech recognition, but no examples of people using the Kinect's RGB camera for contour-mapping in an effort to perform lip reading. Upon discovering this, we undertook to discover if the Kinect were capable of using the color camera to take in data that could be subjected to further statistical analysis.

Previous approaches to the problem of optically enhanced speech recognition have used a wide variety of tools. We've come across reports[7, 12, 8] using general-purpose cameras and/or custom-made infrared rigs. More recent reports utilize the depth-finding features of the Kinect, both alone and in parallel with another camera.

3. METHODOLOGY

We propose to examine the interface Microsoft provides for the Kinect; we attempt to quantify the usability of the tools they provide us, with an eye to the granularity of their results. We hypothesize that we can access complete high-resolution models of the lips through the color camera, recognition of specific syllables and allophones through the microphone array, and a completely functioning link between these two separate systems. In the event that we find support for these operations, we believe further research should be carried out with the goal of implementing speech recognition via the Microsoft Kinect.

In order to ascertain whether computerized AVASR was possible, we needed to examine the range and accuracy of audio and visual data offered by the Kinect. As purely audio-based speech recognition is known to be offered in some form by the Kinect, the audio test's primary function was to examine the interface offered, and determine the accuracy with which the recognition context identifies words in different conditions. While face-tracking is known to be offered by the Kinect, sufficient granularity for lip-tracking and analysis of mouth shapes was unknown. Therefore, the test of the visual abilities of the Kinect was focused on identifying how much data could be extracted from the RGB video feed.

The following tests were carried out on a laptop PC running Windows 7, with a 2.71 GHz 8-core i7 Sandy Bridge processor and an NVIDIA GeForce GTX 560M video card. We utilized Microsoft Visual Studio as our development environment, and all code samples and library functions integrated into our programs were provided by the Microsoft Developer Network, the Kinect SDK, and the Microsoft Speech libraries.

3.1 Audio Testing

We tested the audio functionality of the Kinect by constructing a program that identifies spoken words and reports the confidence with which it had identified them. We used this program to test comprehension on a corpus of predefined words, in a series of different auditory environments, with varying speed and cadence. Testing occurred in two rooms with ambient noise levels measured at 45dB and 75dB, respectively.

The back end of the program is composed of a SAPI audio capture engine. This engine operates in a multithreaded event-driven fashion, wherein a thread of the engine may at any point register a word recognition event. This event contains the context from which it was drawn, including a timestamp suitable for use in linking the audio event to an external source of visual data. The audio delimited by the event is subsequently compared to a user-defined dictionary of matchable words, and the engine returns the most probable match from that grammar. The percentage-certainty is recorded and passed along with the other context of the word or phrase.

Word	Certainty at 75dB	Certainty at 45dB
Giraffe	77%	93%
Example	75%	80%
Maybe	78%	95%
Font	73%	83%
Parallel	73%	78%
Corpus	79%	93%

Table 1: Audio results

3.2 Video Testing

The testing process for the Kinect was designed to answer two related questions: Does the Kinect API offer sufficient capacity to record lip movements with some degree of detail? If so, under what parameters are those lip movements sufficiently recognizable to allow for further analysis? These two questions can be reduced to answering whether, for this purpose, the Kinect is both sufficiently usable and sufficiently accurate.

The usability of the Kinect was measured by endeavoring to develop a small program to somehow record lip movements in some quantifiable way. As the Kinect is known to be capable of recognizing the presence of faces, this was then a question as to the granularity of that recognition system. A sufficiently robust interface would allow for isolating the lips from the rest of the face, and for saving those data points in such a way as to preserve enough context to allow for its usage in tandem with the audio speech recognition facilities of the Kinect.

The visual acuity of the Kinect was tested by analyzing the performance of the above program by introducing a series of variables to the process. In an effort to simulate a visually "noisy" room, we sat the subject in front of the Kinect and set several other individuals to pacing about behind the subject at varying distances and angles. By varying the distance between the subject and the sensor, we attempted to establish optimal ranges on the operation of the recognition program. The subject was instructed to move his head and torso in and out of the frame, and introduce changes in the pitch, yaw, and roll of the head, in an effort to discover at which angles the lip reading process may be degraded.

4. EXPERIMENTAL RESULTS

4.1 Audio results

The following results were had by leveraging Microsoft developer resources and code samples provided with the Kinect SDK. In the course of our research, we show that speech capture and recognition of a set of randomly-selected English words based solely on audio data is possible with the Kinect.

Table 1 shows an comparison of a series of speech recognition events at different ambient noise levels. Five words were selected as representative examples from the grammar, and the corpus as a whole is represented as well. The percentages displayed in each column represent the average certainty values from five separate tests carried out sequentially.

While the recorded certainty varies between 73% and 95%, best results (>90%) for the entire grammar of words were had when a single speaker sat within a meter of the Kinect, with ambient noise less than or equal to 45dB. Recognition certainty percentages dropped linearly with distance from the recorder, and failed recognition of major segments of the corpus occurred only in speech intentionally garbled or not delimited by pauses. With increased noise, there was an increased incidence of false starts, which is to say, recognition events that registered words from the background noise, with less than 50% certainty ratings.

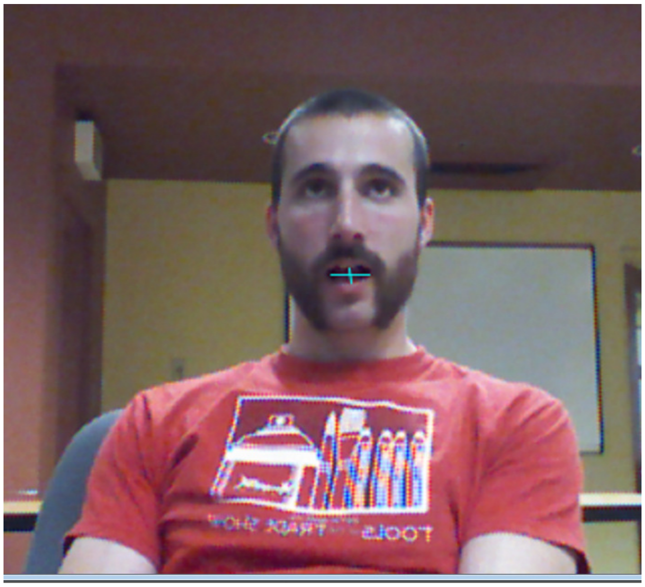

Figure 3: An early model of our lip-tracking model. Further improvements have allowed modeling of 22 individual vertices within the mouth environment.

We also observed that the Microsoft Speech library supports access to timestamps within the audio stream for recognition events, allowing us to tie word data to specific moments in the audio stream, for subsequent synchronization with the video stream. Furthermore, the interface exposes hypothesized words during frequent stages of the recognition process. This would allow future researchers greater access to the decision-making process, in order to better inform their statistical models.

4.2 Video Results

The program we used to test the visual acuity of the Kinect, which leveraged the Face Tracking API, proved capable of representing the face as a series of 100 vertices, describing positions along an X/Y axis delimited by the edge of the camera's field of view. The lips are represented by an array of 22 data points associated with the context of the subject's face. This recognition took place in real-time, and allowed us to extract a high-resolution representation of the movement of the lips. The Kinect captures the lip movements at a rate that varies between 9Hz and 30Hz, inversely proportional to the resolution of the image. The RGB camera of the Kinect operates at a default resolution of 640x480, but is capable of capturing images with a resolution of 1280x1024 at the lower end of the frame rate scale.

The camera's capability for lip-tracking in the face of various alterations to the environment was encouraging. While the Kinect's rangefinder is capable of automatic recalibration, the RGB camera advertises optimal recognition beyond 3 feet. We found that, for the purposes of contour-mapping, the optimal range for lip recognition lies between 60 cm and

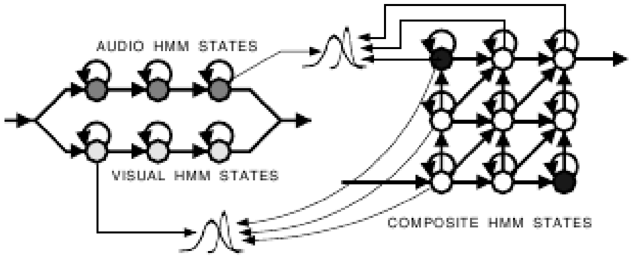

Figure 4: Hidden Markov Model relating audio and visual states. This type of modeling is a promising approach for correlating audio data from the Kinect with captured video.

1 meter.

As the distance between subject and sensor was decreased below 60 cm, the facial mesh became progressively less well-defined, flickering rapidly between states as it attempted to lock down the face, and, by extension, the lips. This stuttering effect introduced noise to the data, which, up to a point, could be filtered out by monitoring the data for unrealistically rapid changes in mouth shape. As the distance between subject and Kinect was increased beyond 1 meter, the vertices representing the lips became less well-defined as the camera lost the resolution necessary to identify the fine lines of the mouth. By decreasing the frame rate, we were able to extend the range of optimal accuracy to approximately 2 meters.

In the zone between 60 cm and 1 meter, we undertook to examine the responses of the facial mesh to tilts and swivels of the subject's head. We found, unsurprisingly, that the accuracy of the lip model was greatest when the subject was facing directly towards the camera. The lip models deformed fairly accurately as the subject began to turn, and only lost cohesion completely when an edge of the mouth became no longer visible. Nevertheless, the image of the lips in profile (however accurate) did present a problem. As the angle at which the lips were observed increased, the level of data offered by that lip model suffered a corresponding decrease. By viewing the mouth in profile, the Kinect was unable to capture as large of a variation in movement over spoken words, decreasing the ease with which individual uttered allophones could be identified.

5. TOWARDS LIP READING WITH KINECT

We posit that we are able to gain access to enough data to warrant further research into the realm of contour-mapping-based lip reading using the Microsoft Kinect. There are many angles on which we could focus in subsequent research of our problem. Most promising is the furthering of the lip-tracking, which was mostly theoretical in the course of our experiments. By addressing the issue with a focus on the integration of audio- and video-based recognition techniques, we have a much wider variety of metrics to analyze. Our plan of action, going forward, is to synchronize lip movement with words that have been recognized through the Kinect API. This can be made possible through support for timecoding the audio and video stream, which is already present. With this information, we can take the series of generated lip models and correlate them with the sound associated with speaking a word.

Should two-dimensional representations prove insufficient for the necessary operations, it is possible to utilize the infrared depth finder to create millimeter-accuracy[9] three-dimensional maps of the subject's teeth and mouth. These additional data points make it feasible to detect a wider variety of the allophones constituting human speech. Previously published efforts at speech recognition with the Kinect have only had access to depth models in their efforts to map the lips, so we would be well-placed to merge our results with existing academic data. A particularly promising matching method involves a Hidden Markov Model, as used in the audio analysis program used by Potamianos et al. [11], and shown in Figure 4. Another potential approach is using machine learning classification methods such as support vector machines, wherein features are represented by the lip models that comprise the sounding of a word. Such an approach could help us to distinguish various allophones and to then generate the corresponding text. Allophones themselves can be correlated to the phoneme that they generate, and based on rules of English, the phonemes can be classified and corrected to form words. White et al., considered this approach for reconstructing encrypted VoIP conversations [14], and such an approach would be certainly suitable to lip reading, where the linguistics community has studied the relationship between lip models and the allophones and phonemes produced [6, 1]. Having the timecoded audio provides a ground truth that we can use to refine our models, with the eventual goal being the removal of the audio track and examining whether phoneme and word reconstruction is still possible. We postulate that this course of action is plausible.

Should we reach a situation in which our lip reading produces sufficiently satisfactory results, we could implement a COM-compatible lip reading engine. We have found that SAPI provides an interface for alternative speech recognition engines, allowing us to build our own visual speech recognition engine and attach it to SAPI, thereby producing a two-way feedback mechanism for the further improvement of the link between the algorithms. We can then investigate whether phoneme reconstruction is possible with lower granularities of lip modeling, e.g., when a subject is farther away from the Kinect and partially turned from the camera, or when there are multiple subjects in the room to be tracked.

What this work has shown is that, while many have prognosticated on the ability of future versions of the Kinect to perform lip reading, our preliminary results and potential work forward demonstrate that existing deployed devices already possess this potential. The Kinect is a particularly well-designed device for this functionality, given the multiple cameras and rangefinding ability, and provides an excellent reference point for this work; however, the techniques that we devise will have the potential for working with any device that captures video input, including webcams and smartphones. Already work such as PlaceRaider [13] has shown that smartphones infected with "visual malware" can be used to generate a three-dimensional reconstruction of an indoor space. By being able to capture lip reading and reconstruct phrases and sentences, these devices surpass the expectations that users have of their abilities and open the door for major future concerns about what can be spoken privately. May the future hold the ability of our devices to act as the HAL 9000 from "2001: A Space Odyssey", and spy on our conversations while we are oblivious to the true capabilities of our household devices? We intend to explore these possibilities.

6. ACKNOWLEDGMENTS

We thank Microsoft Research for providing support for this work through Project HAWAII. We also thank Ellen Klowden, MSW, for editorial assistance.

7. REFERENCES

[1] J. D. Amerman, R. Daniloff, and K. L. Moll. Lip and Jaw Coarticulation for the Phoneme /ae/. *Journal of Speech and Hearing Research*, 13(1):147, Mar. 1970.

[2] J. Biggs. iFixIt's Kinect Teardown, November 2010.

[3] C.-Y. Chang, B. Lange, M. Zhang, S. Koenig, P. Requejo, N. Somboon, A. A. Sawchuk, and A. Rizzo. Towards Pervasive Physical Rehabilitation Using Microsoft Kinect. In *The 6th International Conference on Pervasive Computing Technologies for Healthcare*, San Diego, CA, May 2012.

[4] R. A. Clark, Y.-H. Pua, K. Fortin, C. Ritchie, K. E. Webster, L. Denehy, and A. L. Bryant. Validity of the Microsoft Kinect for assessment of postural control. *Gait & Posture*, 36(3):372–377, 2012.

[5] J. Condliffe. Kinect 2: So Accurate It Can Read Lips? Gizmodo: `http://gizmodo.com/5862968/kinect-2-so-accurate-it-can-lip-read`, Nov. 2011.

[6] R. Daniloff and K. Moll. Coarticulation of Lip Rounding. *Journal of Speech and Hearing Research*, 11(4):707, Dec. 1968.

[7] G. Galatas, G. Potamianos, D. Kosmopoulos, C. McMurrough, and F. Makedon. Bilingual corpus for AVASR using multiple sensors and depth information. In *Proc. of AVSP*, pages 103–106, 2011.

[8] J. Huang, G. Potamianos, J. Connell, and C. Neti. Audio-visual speech recognition using an infrared headset. *Speech Communication*, 44(1):83–96, 2004.

[9] K. Khoshelham. Accuracy analysis of Kinect depth data. In *ISPRS workshop laser scanning*, volume 38, page 1, 2011.

[10] B. F. Mentiplay, R. A. Clark, A. Mullins, A. L. Bryant, S. Bartold, and K. Paterson. Reliability and validity of the Microsoft Kinect for evaluating static foot posture. *Journal of Foot and Ankle Research*, 6(14), Apr. 2013.

[11] G. Potamianos, C. Neti, J. Luettin, and I. Matthews. Audio-visual automatic speech recognition: An overview. *Issues in Visual and Audio-Visual Speech Processing*, 22:23, 2004.

[12] M. Rafati, M. Yazdi, A. Seyfi, and M. Asadi. Realtime Lip Contour Tracking For Audio-Visual Speech Recognition Applications. *International Journal of Biological and Life Science*, 4(4):190–194, 2008.

[13] R. Templeman, Z. Rahman, D. Crandall, and A. Kapadia. PlaceRaider: Virtual Theft in Physical Spaces with Smartphones. In *NDSS'13: Proceedings of the 20th ISOC Symposium on Network and Distributed Systems Security*, San Diego, CA, USA, Feb. 2013.

[14] A. M. White, A. R. Matthews, K. Z. Snow, and F. Monrose. Phonotactic reconstruction of encrypted VoIP conversations: Hookt on fon-iks. In *Proceedings of the 32nd IEEE Symposium on Security and Privacy*, pages 3–18. IEEE, May 2011.