A LOW COST NON-CONTACT APPROACH TO TONGUE TRACKING FOR SPECIAL NEEDS CHILDREN

Kyle Mulligan, Jonathan LaRocque, and James Green
Department of Systems and Computer Engineering, Carleton University, Ottawa, Ontario

INTRODUCTION

Children who suffer from speech difficulties are often faced with long speech therapy regimes for which there is little motivation for the child. The goal of this project is to develop an interactive computer system that will track a child's mouth and tongue and guide them through a series of exercises disguised as a game. An effective solution may provide sweeping benefits for these children as many speech difficulties are due in part to poor tongue motor skills. For example, lalling, involving misarticulated 'r', 'l', 't', and 'd' sounds, may be caused by poor control of the tongue tip [1].

Ultimately, the system will consist of two principle components: 1) a safe non-invasive tongue tracking system, based on an inexpensive webcam; and 2) an interactive program or game that will engage the child and lead the child through a series of exercises. A prototype solution for the first component is described here. The prototype tongue tracking system makes use of a generic low-cost webcam that tracks the state of the user's tongue and mouth. Although being limited to tracking only a protruding tongue, this low-cost non-contact approach offers significant advantages over some current systems that make use of specialized ultrasound equipment [3] or invasive sensors placed in the mouth [2] which is not a viable solution for child users. The emerging use of webcams and availability of inexpensive desktop computers increases the feasibility of a low cost speech therapy device based on these off the shelf components.

The state of the user's mouth (open/closed) and tongue (left, right, up, down, in, out) is interpreted by the computer vision system and can be translated into control inputs for a fun and interactive game. By providing the child user with such an interface, we hope to make speech therapy seem less like a job and more like a fun activity. It is hoped that usage of this system will lead to increased positional awareness, strength, and control of the tongue and mouth for the user. While originally targeted to children with Down syndrome, this system is widely applicable to other children undergoing speech therapy.

METHODS

There are many challenges in developing image processing software to locate and track a user's tongue and mouth. The major steps are:

1. Determining the minimum camera performance specifications with the goal of a modular design;
2. Developing a robust, lighting independent algorithm for automated mouth identification;
3. Developing a rapid face tracking algorithm to compensate for user movements;
4. Determining the state of the mouth (open/closed) and the direction of the protruding tongue.

The overall procedure of the tongue tracking system is shown in Figure 1 below.

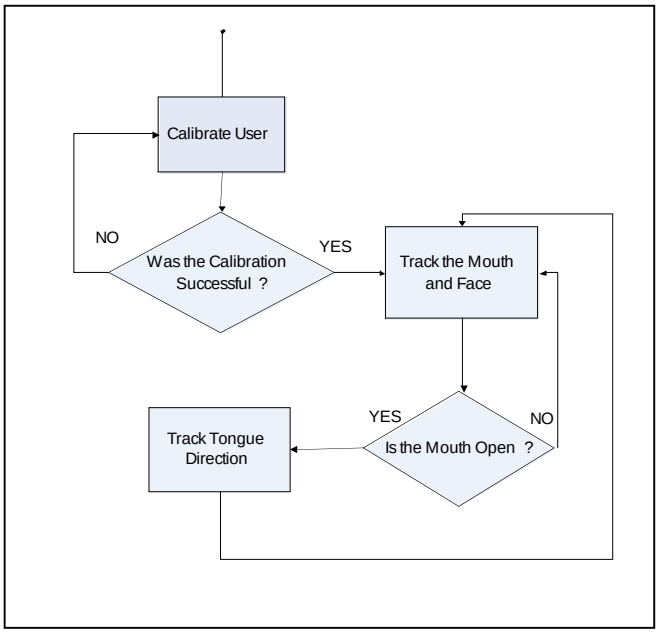

Figure 1: Tongue tracking algorithm flow chart

A. Pre-processing of the image

One of the goals of this project is to produce a modular design. Therefore it was important to dynamically adjust the image brightness and contrast

to remove dependencies on particular camera systems or lighting conditions. An algorithm was developed to stretch each channel (R=red, G=green, B=blue) to use the full range (0-255) in each frame. This algorithm was required due to the wide range of colour interpretations of the web cameras tested in this study. The algorithm reduces systematic bias between camera systems. This step is critical since the mouth detection and tongue tracking algorithms are both based on searching for pixels within experimentally determined RGB intensity ranges.

B. Automated calibration: Locating the mouth

Prior to tracking the state of the mouth and tongue, the initial position of the mouth must be known. A robust automated and non-contact calibration method is preferred if this system is to be widely adopted in non-expert family settings and is to be used with children.

Automated calibration begins by requiring the user place his/her face within a bounding box overlaid on their real-time moving image on the computer screen. The user is asked to open his/her mouth and click on a button. At this point, the automated mouth segmentation algorithm (described below) will automatically locate the mouth and draw a "mouth box" around the individual's mouth (see Figure 2). If the auto-calibration is successful, the user is asked to close his or her mouth and then click on "Calibration Successful" button. As discussed below, this image of the closed mouth is later used to determine the state of the mouth (i.e. open vs. closed). If the auto-calibration fails, the user then has the choice to either repeat auto-calibration or to initiate manual calibration. During manual calibration, the user is instructed to place his/her mouth within a tight bounding box and to click a button when he/she is ready.

The automated mouth segmentation algorithm examines all pixels within the bounding box to identify those pixels that lie within experimentally derived RGB thresholds (see Results). To estimate the vertical centre of the mouth, all pairs of pixels within each column are examined for proper vertical separation. The range of allowable separation is proportional to the size of the initial bounding box that the user was asked to fill with his/her face, and makes use of observed anthropometric data relating the expected size of the mouth to the size of the head. The average vertical position of all qualifying pixel pairs is used as the vertical center of the mouth. Likewise, each row is examined to identify the horizontal centre of the mouth. This algorithm proved to be highly robust despite variations in lighting conditions and skin tone (see Results).

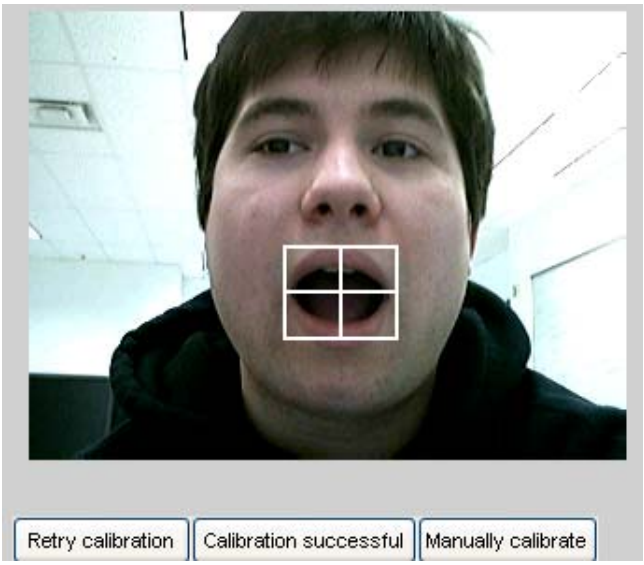

Figure 2: Result of automated calibration phase

B. Tracking the mouth

Considering that this device is aimed at child users, a significant degree of user head motion is expected. Unlike ultrasound-based systems that require the user's head to be perfectly still necessitating use of a helmet [4] or a chin rest [3], the present system unobtrusively tracks the user's face and mouth, allowing increased freedom. Once the initial mouth position is determined through the methods described above, the mouth tracking algorithm is used to compensate for natural movement of the user (within the bounds of the webcam) without adversely affecting the tongue tracking. Mouth tracking is performed using standard image correlation functions which compare an original image to a potentially shifted image from a later frame. The information obtained from this correlation provides the relative offset of the mouth for each frame compared to its position in the original image. The bounding box of the mouth is then moved accordingly.

C. Determining the state of the mouth and tongue

The state of the mouth (i.e. open vs. closed) is assessed in each frame via a template matching algorithm. The contents of the mouth box within each frame are compared to the 'mouth closed' image recorded during calibration. A decision threshold is applied to classify the mouth as open or closed. Template matching was an ideal method in this case since there was an opportunity to take a snapshot of the user's closed mouth immediately following calibration. For a relatively simple 2-state prediction problem, such as mouth open vs. mouth closed, only a single training image of the mouth closed is required.

Therefore the calibration/training phase is not overly onerous for the user.

If the mouth is determined to be open, tongue tracking ensues. Two methods were considered for the actual tracking of the tongue. One involves the use of template matching, and the other involves using knowledge of colour values of the tongue and surrounding tissues (i.e. lips, teeth and skin around the mouth). In both cases, processing must be done on each captured frame to determine the position of the tongue.

The template matching method would require the user to perform a prolonged calibration phase, in which multiple snapshots of each possible position of the tongue would be taken and stored as training data. These snapshots would later be used to determine the position of the tongue by comparing them (by correlation) to the observed mouth image in the current frame. Also, due to the fact that correlation is CPU intensive, an issue of performance may arise since many comparisons would be done for each frame captured by the webcam. Therefore, the second option was pursued for tongue tracking.

In order to track the protruding tongue, the "tongue colour value" method is used. In order to select the range of R, G, and B values that will correspond to tongue tissue, a number of individuals with various ethnicities (and thus varying skin tones) were asked to sit in front of the webcam to allow for the measurement of the colour of their tongues (measured in RGB intensities). This information was used to create a range of accepted tongue colour values (see Results).

During real-time tongue tracking, the sub-image of the mouth (bounded by the "mouth box") is extracted from each captured frame. All pixels whose RGB values fall within the predefined red, green and blue ranges are considered. For each of four possible directions, an 'ideal mask' is then applied to the image and a sum of matching pixels is computed. The current state of the tongue is taken to be the direction which has the highest evidence score. See Figure 3 for an illustration of the application of an ideal mask for the 'up' direction. Note that pixels from the lips will often fall within the tongue RGB range. This does not adversely affect tongue tracking performance since this will bias each of the four tongue directions equally and we take the maximum score.

Although they are experimental, the current ranges of RGB values have yielded promising results which strengthen the justification for choosing this method over the template matching method.

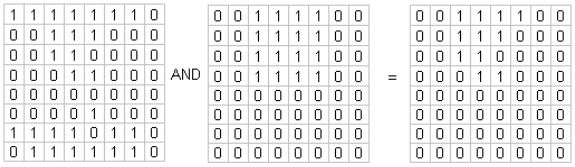

Figure 3: Illustration of tongue tracking algorithm

RESULTS

Table 1 shows the measured lip and tongue pixel intensity value ranges in the red, green, and blue channels obtained from group of individuals with different ethnicities and skin tones (Caucasian, Egyptian, African, and Latino). Note that the value of the red channel provides the greatest discriminant power between the lips and tongue.

Table 1: Lip and tongue pixel intensity ranges for various skin tones

	Colour channel	Lower bound	Upper bound
LIP	R	80	100
	G	40	80
	B	40	80
TONGUE	R	110	140
	G	65	85
	B	50	85

Figure 4 shows the pixels on a user's face that match the lip RGB ranges from Table 1 above. As can be seen, the entire lip area has been identified as well as a number of other regions of the face. The algorithm described above makes use of known anthropometric proportions to correctly identify the mouth. An experiment was conducted to evaluate the accuracy of the auto-calibration system. Eight subjects of varying ethnicity and skin tone (4 Caucasian, 2 east-Indian, 1 Persian, 1 Latino) attempted the auto-calibration stage in a computer lab with ambient lighting conditions. The mouth was correctly identified for all subjects with the exception of subject 5 (Caucasian) and subject 7 (east Indian). Auto-calibration was successful when repeated for subject 5. Subject 7 required manual calibration (no red pigmentation in lips).

To evaluate the face tracking algorithm, the same eight subjects were then asked to move within the bounds of the camera field of view. All faces were successfully tracked.

Figure 4: Pixel regions matching RGB ranges specified in Table 1 are highlighted on this image

Four subjects were then used to evaluate the mouth state detection algorithm (i.e. mouth open vs. closed). Each subject was asked to open and close their mouth ten times and the predicted state was recorded. The overall accuracy was 90%. The system was highly reliable when the tongue was within the mouth. A protruding tongue was occasionally mistaken for a closed mouth due to the similarity in RGB values for tongue and lip pixels.

Finally, the tongue tracking algorithm was tested on four subjects (2 Caucasian, 1 Egyptian, 1 Latino) using four possible states: up, down, left, and right. Overall, the 4-state accuracy was 90%. The majority of errors were due to confusion between left vs. right, whereas up vs. down was highly accurate. A narrower 'ideal mask' for the left and right directions should improve accuracy by excluding more of the lip pixels. Note that all four test subjects had lighter skin tones. The system is expected to perform even better for users with darker skin as this will provide a higher contrast between the lips and the tongue.

DISCUSSION

A. Lighting issues in dark environments

Even though an image pre-processing algorithm was implemented to improve and standardize the quality of images acquired using a range of low to high-end web cameras, we observed that in environments with low lighting the automatic calibration stage occasionally failed to detect the mouth. The manual calibration proved useful in such situations. Also, the introduction of inexpensive USB powered lights has led to a significant improvement to the auto calibration accuracy in these environments.

B. Future Work

While work will continue on improving the robustness and accuracy of the tongue tracking system, the prototype system is largely in place. The second principle component of this system will be involve the creation of fun and interactive games in order to encourage child users to engage the system and strengthen their tongues. Research into intuitive interfaces will be conducted. For example, early user tests of the current system led to the adoption of a mirror image (i.e. horizontally flipped) interface to lessen the training time required for new users. A number of games that appeal to children in various age ranges will be developed based on the 5 state control inputs from the tongue tracking system (i.e. mouth closed, tongue left, up, down, and right). Clinical research will then be conducted to determine if use of this system leads to improved motor skills of the tongue. Through collaboration with a speech therapy research center, we will evaluate this system with children undergoing speech therapy to measure user acceptance and benefits in terms of tongue strength and control and also resulting speech quality.

CONCLUSIONS

A novel approach has been described to improve the quality of life of children with speech difficulties. This system attempts to improve and reinforce the fundamental motor skills of the tongue in the hopes of improving speech quality. A prototype tongue tracking system has been developed and has been shown to achieve high reliability. The next step is to develop games to encourage children undergoing speech therapy to exercise their tongues. Through collaboration with clinical researchers, we will determine if the device improves the motor skills of the tongue yielding improved speech quality. We hope to show that this system is a beneficial addition to current speech therapy regimes. An equivalent system to help other speech therapy patients such as stroke victims will also be explored.

REFERENCES

[1] "Speech Disorder." McGraw-Hill Encyclopedia of Science and Technology. McGraw-Hill, 2005.

[2] M. Aron, M. Berger, E Kerrien, and Y. Laprie, "Coupling electromagnetic sensors and ultrasound images for tongue tracking," *7th International Seminar on Speech Production,* Ubatuba, Brazil, Dec. 13-15, 2006.

[3] M. Stone, "A guide to analyzing tongue motion from ultrasound images," *Clinical Linguistics and Phonetics*, vol. 19, pp. 455-502, 2005.

[4] J. Scobbie, K. Sebregts, and J. Stuart-Smith, "From subtle to gross variation: an Ultrasound Tongue Imaging study of Dutch and Scottish English /r/", *Tenth Conference on Laboratory Phonology*, Paris, June 29-July 1, 2006.