MAPPING A CONTINUOUS VOWEL SPACE TO HAND GESTURES

Yadong Liu[*1], **Pramit Saha**[†2], **Arian Shamei**[‡1], **Bryan Gick**[♦1,3], **and Sidney Fels**[#1]
[1]Dept. of Linguistics, University of British Columbia, Vancouver, British Columbia, Canada
[2]Dept. of Electrical Computer Engineering, University of British Columbia, Vancouver, British Columbia, Canada
[3]Haskins Laboratories, New Haven, Connecticut, USA

1 Introduction

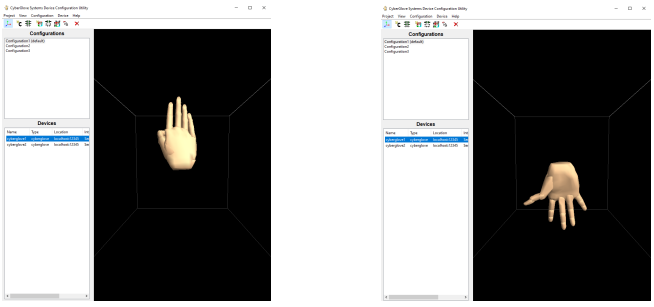
Individuals with speaking disabilities often use Text-To-Speech (TTS) synthesizers for communication. However, users of TTS synthesizers often produce monotonous speech and the use of such synthesizers often renders lively communication difficult [1]. As a result, hand gestures have been used to successfully generate of speech [1, 2]. Fels and Hinton [2] designed Glove-TalkII that translates hand gestures to spoken English via an adaptive interface. The system allows users to generate an unlimited number of English vocabularies by controlling ten parameters of the speech synthesizer [2]. Each parameter maps to a different hand gesture or location, allowing the user's hands to act as an artificial vocal tract [2]. Another hand-gesture-to-sound mapping system developed by Kunikoshi et al. [1] maps a set of five hand gestures to the five vowels of Japanese with smooth transitions. However, both systems have their limitations. The first system uses extensive hand movements and is less intuitive for an interested layman. The second system is designed to synthesize speech sounds of only one language and uses distinct hand gestures to represent individual speech sounds, as a result of which, continuous change between two speech sounds cannot be intuitively represented by continuous hand movement.

The goal of this project is to develop a synthesizer for which hand movement can be used to control and produce a continuous vowel space more easily and intuitively. It also aims to make the synthesizer more user-friendly by providing real-time speech sounds as feedback, as well as an inverse model that visualizes hand gestures which are required for specific sounds.

2 Proposed method

2.1 Data collection

We use CyberGlove II, manufactured by Immersion Inc., to capture hand movements. It has 18 sensors that record information such as wrist flexion and bend and abduction of the fingers. This two-dimensional (2D) control of flexion and abduction allows the user to control the 2D formant space continuously. Figure 1 shows the hand gestures for [a] (fingers adducted, pointing upward) and [i] (fingers abducted, pointing downward). The elbow is fixed at rest position and utmost care is taken in order to avoid all other unintended motions.



(a) Hand gesture for /a/ **(b)** Hand gesture for /i/

Figure 1: The Virtual Hand SDK of Cyberglove

2.2 Forward and inverse mapping between kinematics and acoustics

Wrist flexion and abduction of the index and pinky fingers are used to mapped linearly onto F1 and F2 respectively as shown below, which are formant values inherent to vowel sounds. Speech sounds are then synthesized using Vowel Synthesis in MATLAB [3].

$$F1 = \frac{F1\max - F1\min}{flexion\max - flexion\min} \times flexion + b1$$
$$b1 = F1\min - \frac{F1\max - F1\min}{flexion\max - flexion\min} \times flexion$$
$$F2 = \frac{F2\max - F2\min}{abduction\max - abduction\min} \times abduction + b2$$
$$b2 = F1\min - \frac{F2\max - F2\min}{abduction\max - abduction\min} \times abduction$$

A male adult user performs hand movements such as adducting and abducting four fingers together with bending the wrist up and down. Estimated F1 and F2 from hand movements are used as input in a vowel synthesizer to generate vowels. F0 is fixed as 100 Hz, and F3 is fixed as 2400 Hz.

In addition to the pilot results, more hand movements were mapped to different vowels. In order to facilitate the learning process of using CyberGlove to produce vowels, inverse modeling was designed to visualize hand movements needed to produce certain sounds. F1 and F2 values from a sequence of vowels of spoken English was extracted using FormantPro [4] and then converted to wrist flexion and finger abduction. The wrist flexion and figure abduction that being generated by users hand can be compared to those converted from inverse modeling. Thus, users is informed to perform in a more accurate way. Based on the linear regression analysis, the value of $b1$ and $b2$ were fixed at 482.64 and -332.79 respectively.

$$flexion = (F1 - b1) \times \frac{flexion\max - flexion\min}{F1\max - F1\min}$$
$$abduction = (F2 - b2) \times \frac{abduction\max - abduction\min}{F2\max - F2\min}$$

[*]yadong.liu@ubc.ca (equal contribution)
[†]pramit@ece.ubc.ca (equal contribution)
[‡]arian.shamei@ubc.ca
[♦]gick@mail.ubc.ca
[#]ssfels@ece.ubc.ca

3 Experiments and results

The first and second formant frequencies used for synthesizing the outer cardinal vowels are depicted in Table 1.

Table 1: Formant frequencies

Vowels	F1	F2
[i]	270	2290
[ɑ]	730	1090
[æ]	660	1720
[u]	300	870

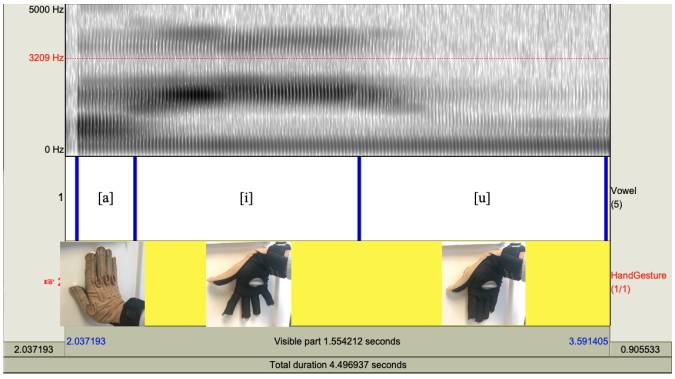


Figure 2: The spectrogram corresponding to [a], [i], [u] and respective hand gestures

Figure 2 shows the spectrogram of three vowels with smooth transition in between generated from vowel synthesizer with continuous hand movement as input. First layer in the figure presents the spectrogram, second layer shows segmentation, and third layer presents corresponding hand gesture.

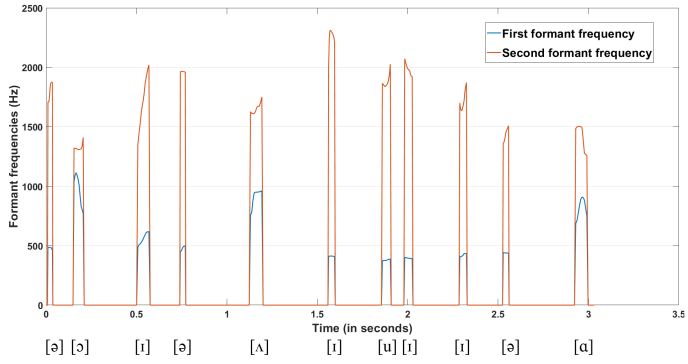


Figure 3: First and second formant frequency corresponding the vowels of the sentence "The north wind and the sun were disputing which one was the stronger"

Figure 3 shows F1 and F2 values of vowels in an English sentence "The north wind and the sun were disputing which one was the stronger". Further, those values were converted to wrist flexion and figure abduction as shown in Figure 4. Positive wrist flexion value indicates hand and fingers were pointing upward, and negative wrist flexion value indicates hand and fingers were pointing downward to the ground.

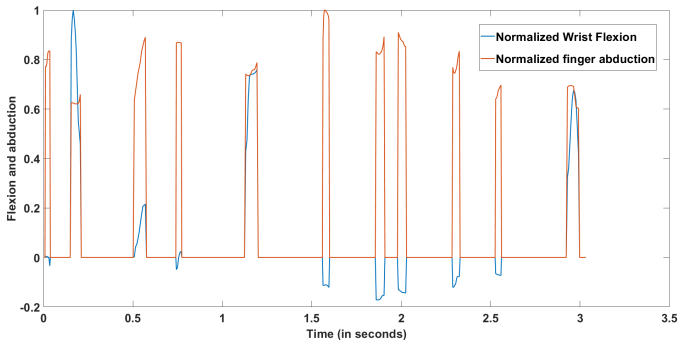


Figure 4: Normalized wrist flexion and finger abduction values for synthesizing vowels of the sentence "The north wind and the sun were disputing which one was the stronger"

4 Discussions and conclusions

This interface primarily uses the CyberGlove as an input device to map continuous hand gestures to English vowels, but, can be easily extended to vowels of any other language. A major advantage of the interface is that its scope is not limited by any particular vocabulary, as it uses formant based vowel synthesis and hence allows synthesis of vowels throughout the vowel quadrilateral, rather than a discrete synthesis of selected vowels. Since this involves a direct one-to-one mapping of the control dimensions to the formant frequencies, it is very easy to learn. Besides, the gestures are intuitive even for a layman without linguistic knowledge, thereby making the interface sufficiently user-friendly. It is also efficient in mapping the gesture transitions accurately to the targeted dipthongs (e.g., in boy), and to a targeted vowel sequence (e.g., in Hawaii). Further step could be to make this hand gesture to speech sound mapping real-time, such that users can produce speech sounds simultaneously with moving hands.

Acknowledgments

This work was funded by the Natural Sciences and Engineering Research Council (NSERC) of Canada and Canadian Institutes for Health Research (CIHR).

References

[1] Aki Kunikoshi, Yu Qiao, Nobuaki Minematsu, and Keikichi Hirose. Speech generation from hand gestures based on space mapping. In *Tenth Annual Conference of the International Speech Communication Association*, 2009.

[2] Sidney S Fels and Geoffrey E Hinton. Glove-talkii-a neural-network interface which maps gestures to parallel formant speech synthesizer controls. *IEEE transactions on neural networks*, 9(1):205–212, 1998.

[3] MathWorks MATLAB. Matlab r2018b. *The MathWorks: Natick, MA, USA*, 2018.

[4] Yi Xu and Hong Gao. Formantpro as a tool for speech analysis and segmentation/formantpro como uma ferramenta para a análise e segmentação da fala. *Revista de Estudos da Linguagem*, 26(4):1435–1454, 2018.

EDITORIAL BOARD - COMITÉ ÉDITORIAL