Opti-Speech: A real-time, 3D visual feedback system for speech training

William Katz[1], *Thomas Campbell*[1], *Jun Wang*[1, 2], *Eric Farrar*[3], *J. Coleman Eubanks*[3], *Arvind Balasubramanian*[4], *Balakrishnan Prabhakaran*[4], *Rob Rennaker*[2, 5]

[1] Callier Center for Communication Disorders
[2]Department of Bioengineering
[3]Arts and Technology Program
[4]Department of Computer Science
University of Texas at Dallas, Dallas, Texas, USA
[5]Vulintus, LLC. Sachse, Texas, USA

wkatz@utdallas.edu, thomas.f.campbell@utdallas.edu, wangjun@utdallas.edu, eric.farrar@utdallas.edu, j.coleman.eubanks@utdallas.edu, arvind@utdallas.edu, bprabhakaran@utdallas.edu, renn@utdallas.edu

Abstract

We describe an interactive 3D system to provide talkers with real-time information concerning their tongue and jaw movements during speech. Speech movement is tracked by a magnetometer system (Wave; NDI, Waterloo, Ontario, Canada). A customized interface allows users to view their current tongue position (represented as an avatar consisting of flesh-point markers and a modeled surface) placed in a synchronously moving, transparent head. Subjects receive augmented visual feedback when tongue sensors achieve the correct place of articulation. Preliminary data obtained for a group of adult talkers suggest this system can be used to reliably provide real-time feedback for American English consonant place of articulation targets. Future studies, including tests with communication disordered subjects, are described.

Index Terms: speech production, visual feedback, speech therapy, electromagnetic articulography

1. Introduction

Although speech information is transmitted primarily in the auditory/acoustic channel, visual information plays an important role in many circumstances. For example, lip reading (and even 'tongue reading') can boost speech identification in noise for healthy listeners [1, 2]. The importance of visual information has been recognized for language learning. Embodied conversational agents (ECAs) have been developed to provide pronunciation training for a variety of applications. Studies have shown that sophisticated talking heads such as BALDI or ARTUR can provide superior learning than that provided by the auditory channel alone [3, 4]. These findings suggest that visual information can be particularly important for processing speech under situations in which acoustic cues are not easily accessed or integrated.

In addition to language tutoring systems, a variety of real-time visual feedback techniques have been explored in recent years. These include technologies based on (1) electropalatography (EPG) to record the contact between the tongue and the palate during speech [5, 6], (2) ultrasound, which provides a midsagittal view of the vocal tract with images of tongue movement and gestures [7, 8], and (3) electromagnetic articulography (EMA), a measurement system that tracks the position of small sensors attached to the speech articulators [9, 10].

Individuals seeking to regain speech and language capabilities following brain damage may benefit from visual information concerning the movement of their articulators. For instance, patients with apraxia of speech (AOS) show gains from real-time information concerning their tongue movement during speech [9-12]. These gains are postulated to occur from preserved feedback capabilities and impaired feed-forward processing in individuals with AOS [12].

Studies also suggest that adult second language (L2) learners may improve their accent by means of real-time visual articulatory feedback. For instance, American English learners of Japanese have shown superior acquisition and retention of the Japanese post-alveolar flap consonant when trained using an electromagnetic articulography (EMA) system, in comparison to a control learning condition [13]. Similarly, Japanese learners of English are reported to demonstrate enhanced acquisition of tense/lax vowel distinctions under EMA visual feedback [14].

An important issue in the development of these technologies is the creation of suitable interactive visualization systems. Relatively little is known about how human subjects can control the tongue under conditions of real-time visual feedback or how different types of displays or interfaces can facilitate these abilities [15, 16]. To address these issues, we have devised *Opti-Speech*, an interactive 3D interface for real-time speech training. With this system a 3D animated tongue avatar provides real-time visual feedback for tongue movement during speech. To our knowledge, this is the first anatomically-based, real-time system that permits the placement of spatial targets inside and outside of the oral cavity for subjects to "hit" (and thereby receive visual augmented feedback for correct productions).

To summarize, our contributions with this system include an interactive, EMA-based, 3D animated visualization system that allows speakers to observe in real time how their tongue is moving as they utter speech sounds. The system uses spatial target zones corresponding to place of articulation which allows the system to give meaningful feedback to the speakers.

In this report, we investigate whether this system can (1) accurately record talkers' lingual place of articulation for alveolar stimuli in a laboratory speech setting, and (2) provide real-time visual information concerning place of articulation that may be useful for speech training applications.

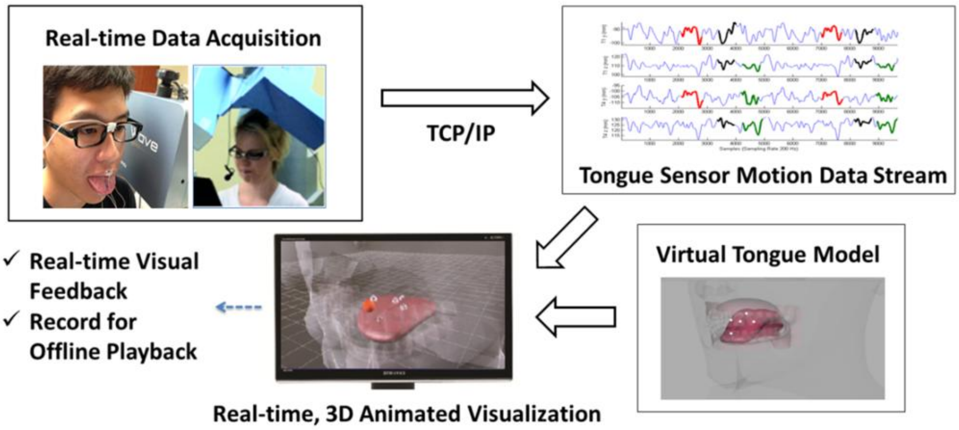
Real-time Data Acquisition
TCP/IP
Tongue Sensor Motion Data Stream
Virtual Tongue Model
✓ Real-time Visual Feedback
✓ Record for Offline Playback
Real-time, 3D Animated Visualization

Figure 1: *Architecture of Opti-Speech system*

2. Description of system

A schematic of the system is shown in Figure 1. Major components include: (i) real-time data acquisition, (ii) tongue modeling, and (iii) visualization. The different modules of the system are described in detail in the following sections:

2.1 Virtual Tongue Model

To visualize tongue motions in real-time, we developed a virtual 3D tongue and human head model using Maya (Autodesk, San Rafael, California). The tongue and head avatar is currently representative of a generic human head and tongue. We eventually plan to incorporate speaker-specific anatomical differences, including palate shape information. The models are set up using standard animation tools so that they can be transformed and animated in a realistic manner.

2.2 Real-time Tongue Motion Data Acquisition

The tongue and human head model are animated based on motion data from the NDI Wave system (Northern Digital Inc., NDI). This technology allows speech motion tracking (in real time) with an error margin of approximately 0.5mm [17]. Tongue movement data (spatial coordinates) are streamed into the visualization component over a network using TCP/IP. Given that we are trying to animate only features most relevant to speech production, and not the entire face (eye position, forehead, etc.) of the avatar, we positioned five sensors from the NDI Wave system on the tongue (Figure 2). A large body of data suggests tongue movement can be described in segments divided into tip/blade and body, with lateral movement being generally symmetrical and particularly important for the production of sounds such as /l/ [18]. A single sensor placed on the chin records the opening and closing of the mouth. A 6DOF reference sensor placed on a pair of glasses tracks gross head movement to avoid skin movement artifacts [19].

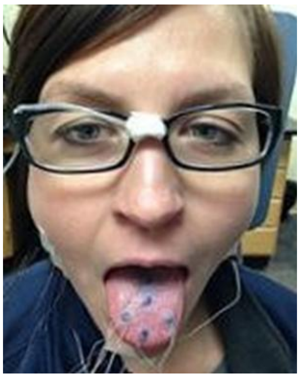

Figure 2: *Positions of tongue sensors and head sensor (on glasses).*

The head-corrected coordinates of the tongue sensors are mapped to virtual control points on the 3D tongue model, which in turn move and deform the tongue mesh to closely match the movement and shape of the subject's actual tongue. The tongue sensor coordinates are "local" to each subject's head, since the data are computed with respect to the reference sensor on the glasses. This allows the virtual head to remain relatively stationary on the screen, allowing the viewer's focus to be on the movement of the tongue within the mouth. The opacity of the head model can also be controlled to allow the tongue to be the most prominent feature of the visualization. This avatar is designed to provide accurate measurements of specific points on the tongue surface and to serve as a visual representation of a tongue, not as a rigorous scientific tongue model.

The system operates in real-time on a standard PC (does not require a high-end graphics card or extensive system memory). To accomplish this we minimized the number of polygons that make up the 3D models. The current tongue and head avatar is made up of approximately 3,500 polygons. For comparison, most video games consist of over 500,000 polygons and are still are able to play in real-time. Our initial results using a standard PC (quad-core 3GHz processor and a 1GB-memory ATI graphics card) show no noticeable time delay or distortion of the tongue model, a predictable output given the small size of our model.

3. Experiment: Operation of real-time target zones

Because some parts of the tongue are relevant for speech and can be easily controlled, it is important to have clearly visible, speech-relevant targets that subjects can "aim for," so that the therapist can help guide the patient with articulatory training. As a result, we have chosen to use the sensor regions as selectable points of reference for guiding tongue shape and motion matching. As a preliminary test of the system, we examined whether productions by healthy talkers can trigger interactive spherical, spatial target zones pre-specified by the experimenters. These regions correspond to place of articulation for English consonants. As a given sensor on the subject's tongue enters the target sphere, the color of the target sphere changes from red to green, indicating a match or a "hit" (Figure 3).

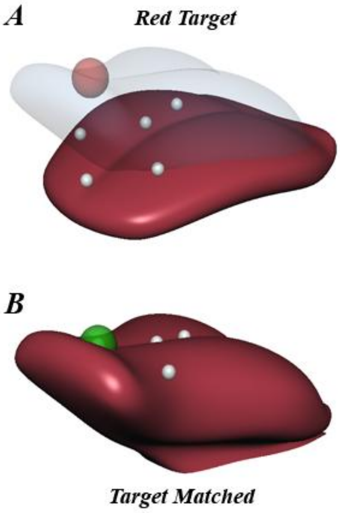

Figure 3: *A) Non-matched tongue-tip target (red). B) Matched tongue-tip target (green).*

3.1. Subjects

Subjects were four native speakers of American English (1 male, 3 female) selected from the University of Texas at Dallas community (ages were 22-32, mean=26). None reported a history of hearing, speech or language impairment. None had prior experience with EMA systems or augmented visual feedback experiments.

3.2. Stimuli

Stimuli were CV syllables containing the vowel /ɑ/ and American English consonants having the following places of articulation: bilabial, interdental, labiovelar, alveolar, post-alveolar, palatal and velar. The consonant /ɹ/ was not included because of its known high degree of articulatory variability. The velar nasal /ŋ/ was not included because it does not occur in syllable-initial position. Thus, the consonants included /p/, /b/, /m/, /f/,/v/, /θ/, /ð/, /s/, /z/, /t/, /d/, /n/, /tʃ/, /t/, /ʃ/, /j/, /k/, /g/, and /h/.

The stimuli were arranged in series of four sequential CV syllables, in which one syllable (randomly) contained an alveolar consonant and the other three syllables began with consonants with different places of articulation, e.g. (/pɑ/-/tɑ/-/gɑ/-/jɑ/; /kɑ/-/ʃɑ/-/mɑ/-/zɑ/, etc.). There were 10 sets of these 4 sequential stimuli in a list, for a total of 40 syllables spoken by each subject.

3.3. Procedure

Each subject was seated in the Wave system facing a computer monitor. Sensors were attached to the subject's tongue tip (1 cm posterior to the apex), tongue middle (~ 2 cm posterior to the tip sensor), tongue body (~4 cm posterior to the tip sensor), and tongue lateral positions. A reference sensor was also attached to a pair of glasses worn by the subject in order to establish a reference for head movement. For this experiment, the tongue tip sensor was used for feedback purposes. The system generated a real-time animation of the subject's tongue within a transparent head model. Investigators next created an alveolar place-of-articulation target zone in this display, based on the subject's perceptually correct production of a series of alveolar CV syllables (/da/, /sa/, or /na/). The target zone is a red sphere which turns to green when the tongue tip sensor enters the space, thus providing augmented visual feedback for place of articulation. Based on previous work [11], we selected a sphere of 1.25 cm in volume (although the target size and "hold time on target" can be varied by the user to make matching easier or harder to guide specific movements). An illustration of the setup is shown in Figure 4.

Two pilot experiments were conducted: a test of the basic fidelity of the system and of a talker's ability to use the device for visual feedback. Subjects #1-3 produced repeated speech without attending to the *Opti-Speech* display. These data addressed the baseline accuracy with which *Opti-Speech* could detect (and denote) place of articulation. This was the primary purpose of this paper. A second experiment (with subject #4) required the participant to attend to the *Opti-Speech* display during speech. These preliminary data further examined visual feedback during speech production.

Following an initial accommodation period to adapt to the sensor (and to allow the subject to produce a range of sounds while viewing the tongue avatar) all subjects were instructed that they would hear series of four nonsense syllables from American English and that they were to then read these syllables aloud. Subjects #1-3 were further informed that although the system would be tracking their tongue movements, it was important to focus on reading the correct syllables (instead of viewing the video monitor). Subject #4 was instructed to watch the video monitor while saying the words and to try and hit the spherical target during the /sa/, /za/, /na/, /la/, /ta/, and /da/ sounds, at which point the target zone would light up green, indicating success.

After two practice trials, subjects repeated the experimental stimulus sets (following the spoken and visual model provided by an investigator). The experiments were video- and audio-recorded.

3.4. Results

Subjects completed the 40-item speaking task without noticeable difficulty (or problems noted at debriefing). Video tapes were later evaluated by one of the authors (WK) and a second independent rater to determine accuracy. Inter-rater reliability was 99%. Table 1 shows the results for the four talkers.

4. Discussion and Future Work

Overall accuracy (hits/total attempts) ranged from 40% to 80% for subjects #1-3 speaking in the no-feedback condition. The system correctly detected most alveolar productions for these subjects. However, the system also registered false positives for subjects #1 and #2, mainly consisting of adjacent (alveolo-palatal and palatal) places of articulation. Subject #3 showed a higher degree of overall accuracy (80%) with no false positives. Although these data are clearly preliminary, two factors seem to be involved in the disparate findings for these three subjects: First, the talkers with lower overall accuracy (and greater dispersion of false positives) exhibited more head

Table 1. *Accuracy patterns for the four talkers. Subject #4 (shaded) received visual feedback.*

Subj (sex)	# Alveolar stim hit	# Alveolar stim missed	# Stim incorrectly hit	Overall Accuracy
1F	10	0	(1) /ʃ/, (1)/j/,(1)/m/, 1(v) (1)/g/,(1) /h/	50%
2F	10	0	(1) /ʃ/, (4) /j/, (1) /g/	40%
3M	8	(1) /n/, (1)/l/	0	80%
4 F	10	0	(1) /ʃ/	90%

movement during speech. At this point in the development of the software, the target region was not sufficiently synchronized to move exactly with the subject's head. This issue has since been resolved. Second, there were potential differences in individual talker anatomy that might explain these results. For instance, subjects #1 and #2 had relatively small vertical tongue excursions, and subject #1 tended to leave the tongue in a high resting position (on the palate) between productions. These factors may have contributed to these subjects hitting an alveolar target region more frequently than subject #3, who showed more vertical tongue movement during speech.

The highest overall accuracy (90%) was shown by subject #4, who received visual feedback for her productions. With the exception of one false positive for the palate-alveolar consonant /ʃ/, this subject accurately hit all alveolar targets.

Taken together, the data suggest that the *Opti-Speech* system can track place of articulation information during speech and allow talkers to access interactive articulatory information. Although the data are clearly limited, the accuracy patterns are rather consistent with predictions of phonetic theory; adjacent place misses (e.g., /ʃ/, /j/) predominate over distal place misses (e.g., /g/, /h/).

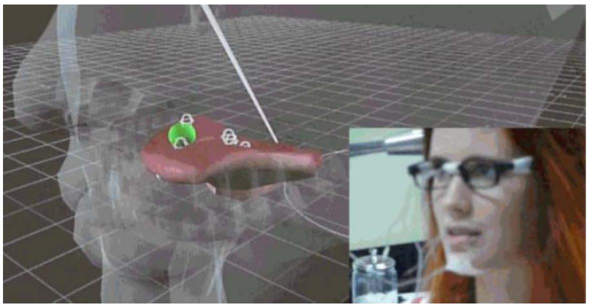

Figure 4: *Illustration of the OptiSpeech system, with subject wearing sensors and head-orientation glasses (lower right insert). Target sphere (green) is being hit during the production of an alveolar consonant.*

http://www.utdallas.edu/~wkatz/alveolar_targets.mov

These data provide some indication of the consistency with which talkers of American English produce repeat tokens of CV syllables and the accuracy with which an EMA interactive feedback system can perform. It further suggests that such a system may be useful for delivering visual feedback concerning place of articulation for accurate consonant production. Similar experiments need to be carried out for vowels (and are currently in progress in our laboratory).

Technical improvements include plans to make *Opti-Speech* operable with input from the AG501 EMA system (Carstens Medezinelektronik, Germany) another commonly-used speech magnetometer system which has slightly higher accuracy (0.3 mm) than the Wave system. We are also exploring software platforms that may provide lower costs and expanded functionality, e.g., Unity (Unity Technologies, San Francisco).

A next important step will be to determine how subjects perform when they have to hit unfamiliar targets, such as foreign speech sounds (or even non-speech gestures). These data are important to obtain for healthy adults and for children learning languages, as well as individuals whose speech and language abilities have been compromised as the result of injury or disease processes.

5. Summary

These preliminary data suggest that the *Opti-speech* system may be used to detect healthy talkers' place of articulation during consonant production under laboratory speech settings. The data further suggest that on-line visual feedback provided by the system may be used by talkers to improve accuracy for articulation during consonant production. However, the present findings are clearly preliminary and more data are needed before firm conclusions can be drawn. Future studies should include tests with more subjects, a greater variety of speech sounds, and conditions involving both laboratory speech and spontaneous speech.

Limitations of the current *Opti-speech* system include the use of a non-individualized representation of the head and vocal tract. Speaker-specific anatomical modeling is being considered for future implementations. Another practical limitation of the current system is the need for a subject to speak with four or five sensors placed on the tongue surface in order to accurately produce realistic motion for the animated model. Although most subjects easily adapt to this sensor array after just a few minutes, it would clearly be less invasive if either fewer sensors were used or if a new, less intrusive interface were designed. We are presently exploring both possibilities.

6. Acknowledgements

The authors gratefully acknowledge support from the University of Texas at Dallas Office of Sponsored Projects, the UTD Callier Center Excellence in Education Fund, and a grant awarded by NIDCD (R43 DC013467). We thank the participants for volunteering their time. We also thank Carstens Medezinelektronik GmbH for material support for this project. We also like to thank Marcus Jones, Amy Berglund, Sonya Mehta, Cameron Watkins, Bill Watts, and Holle Carey for their contribution in data collection, processing, and other support for this project.

7. References

[1] Macleod, A. and Summerfield, Q., "Quantifying the contribution of vision to speech perception in noise", Brit. J. Aud., 21(2):131-141, 1987.

[2] Badin, P., Tarabalka, Y., Elisei, F., and Bailly, G., "Can you 'read' tongue movements? Evaluation of the contribution of tongue display to speech understanding", Speech Communication, 52:493-503, 2010.

[3] Engwall, O.and Balter, O., "Pronunciation feedback from real and virtual language teachers", Computer Assisted Language Learner, 3: 235-262, 2007.

[4] Massaro, D., Bigler, S., Chen, T. Perlman, M., Ouni, S., "Pronunciation training: the role of eye and ear", Proceedings of Interspeech: 2623-2626, 2008.

[5] McAuliffe, M.J. and Cornwell, P.L., "Intervention for lateral /s/ using electropalatography (EPG) and an intensive motor learning approach: a case report", Int. J. Lang. and Comm. Dis., 43(2): 219-229, 2007.

[6] Morgan, A.T., Liegeois, F., and Occomore, L., "Electropalatography treatment for articulation impairment in children with dysarthria post-traumatic brain injury", Brain Injury, 21(11):1183-1193, 2007.

[7] Preston, J.L., Brick, N, Landi, N., "Ultrasound biofeedback treatment for persisting childhood apraxia of speech", Am. J. Sp. Lang. Path. 22(4): 627-643. 2013.

[8] Bernhardt B, Gick B, Bacsfalvi P, Adler-Bock, M., "Ultrasound in speech therapy with adolescents and adults", Clin. Ling. Phon., 9(6-7):605-617, 2005.

[9] Katz, W., McNeil, M., and Garst, D. "Treating apraxia of speech (AOS) with EMA-supplied visual augmented feedback", Aphasiology, 24: 826-837, 2010.

[10] Katz, W., Garst, D., Carter, G., McNeil, M., Fossett, T., Doyle, P. and Szuminsky, N. "Treatment of an individual with aphasia and apraxia of speech using EMA visually-augmented feedback", Brain and Lang., 103:213-214, 2007.

[11] McNeil, M., Katz, W., Fossett, T., Garst, D., Szuminsky, N., Carter, G., and Lim, K., "Effects of on-line augmented kinematic and perceptual feedback on treatment of speech movement in apraxia of speech", Folia Phon. et Logoped., 62:127-133, 2007.

[12] Katz, W.F., and McNeil, M., "Studies of articulatory feedback treatment for apraxia of speech (AOS) based on electromagnetic articulography", Perspectives on Neurophysiology and Neurogenic Speech and Language Disorders, Vol. 20(3):73-80, 2010.

[13] Levitt, J.S. & Katz, W.F., "The effects of EMA-based augmented visual feedback on the English speakers' acquisition of the Japanese flap: a perceptual study", Proc. of Interspeech:1862-1865, 2010.

[14] Suemistu, A., Ito, T., and Tiede, M., "An electromagnetic articulography-based articulatory feedback approach to facilitate second language speech production learning", Proc. of Mtgs. on Acoustics, 19, 060063, 2013.

[15] Shtern, M., Haworth, M.B., Yunusova, Y., Baljko, M., and Faloutsos, P., "A game system for speech rehabilitation", in M. Kallman and K. Bakris [Eds] Motion in Games, Proceedings of the 5th International MiG conference, Rennes, France:43-54, Berlin: Springer, 2012.

[16] Ouni, S. Tongue control and its implication in pronunciation training. Comp, Assisted Lang. Learning, 2013.

[17] Berry, J. "Accuracy of the NDI wave speech research system", J. Sp. Lang. Hear. Res, 54: 1295-1301, 2011.

[18] Ladefoged, P., and Johnson, K. "A course in phonetics, 6th edition", Boston: Thompson Wadsworth, 2010.

[19] Wang, J., Green, J. R., Samal, A., and Yunusova, Y. "Articulatory distinctiveness of vowels and consonants: A data-driven approach," J. Sp. Lang. Hear. Res, 56: 1539-1551, 2013.