A Game System for Speech Rehabilitation

Mark Shtern[1], M. Brandon Haworth[1], Yana Yunusova[2], Melanie Baljko[1], and Petros Faloutsos[1]

[1] York University, Toronto, Ontario, Canada
{mark,mb,pfal}@cse.yorku.ca, brandonh@yorku.ca
[2] University of Toronto, Toronto, Ontario, Canada
yana.yunusova@utoronto.ca

Abstract. Speech disorders are common in children and adults. For a number of these individuals, traditional speech therapy will not be successful, leaving them unable to (or poorly able) to communicate through speech. By combining recent advances in motion tracking technology with state of the art interactive gaming technology, we aim to develop a novel clinical tool that can potentially overcome the shortcomings of traditional methods. There is strong evidence that entertaining and clever visualizations not only can provide necessary augmented feedback to facilitate motor skill acquisition but, also, can be very engaging and effective teaching tools. For the first time, we want to explore the practically infinite possibilities for interactive visualizations of game engines in speech rehabilitation focused on the tongue, the primary articulator, and develop a commercial grade clinical rehabilitation tool with full security, reporting, and data management abilities.

1 Introduction

Speech disorders are relatively common (e.g., 5% of children have a speech disorder in the first grade due to variety of medical conditions; 25% of people who suffer a stroke will have a speech disorder). For 10-15% of these individuals, traditional speech therapy is not successful, leaving them unable (or poorly able) to communicate through speech. To address clinical needs of these populations, new technology-based therapeutic interventions are needed. And even for patients who demonstrate recovery with traditional therapy over time, new interventions can expedite the treatment, minimizing social, emotional and economic burden on the patient, the clinician, and the health care system as a whole.

Speech impairments often arise due to difficulties controlling and coordinating the tongue — the most important speech (and swallowing) organ, but which is hidden. In current practice, clients perform repetitive speech articulation drills under clinician supervision. The clinician judges the client's tongue movements by listening to his or her speech and provides instructive feedback, most often solely based on auditory information, in order to guide the client's tongue placement in subsequent iterations. There are several challenges, however:

– The clinicians may be unable to determine the tongue's movement through listening alone.

M. Kallmann and K. Bekris (Eds.): MIG 2012, LNCS 7660, pp. 43–54, 2012.

- The client may be unable to execute the targeted movement (i.e., in the typical instance, an individual relies on auditory, proprioceptive and tactile feedback during movement execution; however, for individuals with neurological impairment, these feedback mechanisms are often impaired).

There is a mounting body of evidence that augmented visual computer-based feedback techniques can contribute toward speech and language learning outcomes (e.g., see [1] for a summary). However, definitive evidence has been elusive, due, in part, to a host of methodological challenges. Some prior systems have made use of game-like formats, such as Visci et al's *Box of Tricks* [2]. Mechanisms of game play (e.g., the use of rewards and the structuring of difficulty into successive, progressive levels) have been shown to mitigate the negative impacts of demotivation, which often arises in therapeutic domains that required repetitive drills, such as speech therapy. *Serious games*, as they are often called, have been used for a wide range of educational activities ranging from teaching math and physics to high-school children, to training medical personnel, first responders and military units. For example, [3] shows how a computer game can help with social skill development.

In this paper, we are presenting work in progress towards developing a computer game system for speech rehabilitation. We describe a prototype of the system that can track the movement of a patient's tongue in real time and that uses the traces to control on-screen gaming elements whose behaviour relays useful feedback to the patient. Our system, initially, will target adult patients who have experienced stroke, but we hypothesize that the approach is generalizable to other populations of clients who have tongue impairments which result in articulatory and swallowing abnormalities.

Our system integrates two state-of-the-art enabling technologies: the Wave tracking system and the Unity Game Engine. The Wave system — recently developed by Northern Digital Inc., Waterloo, Canada — is capable of tracking a patients tongue and other speech articulators in the oral cavity with submillimeter accuracy [4]. The Unity Game Engine allows us to easily prototype high quality computer games, whose elements can be controlled by the Wave system's motion trackers.

Our system makes use of visual scenes of game scenarios (e.g., a bee flying over flowers), as opposed to those which are based on anatomical models of the vocal tract which have previously been employed (e.g., ARTUR [1], Baldi [5]). We hypothesize that these non-anatomical game-like scenes will evoke a better behavioural response from the users than the scenes that are based on the use of anatomical models, which can be difficult and non-intuitive for non-specialists.

The remainder of the paper is organized as follows. Section 2 describes relevant work in speech therapy and human interaction. Section 3 discusses augmented visual feedback and its benefits to speech therapy. Our framework and its main components are described in Section 4. Initial experiments that showcase the potential of interactive visualization for treatment of speech disorders are presented in Section 5. A discussion of the challenges and future directions are presented in Section 6.

2 Background

The lingual movements during speech are largely hidden, and speakers typically are unaware of precise actions of the tongue during speaking. Immediate and continuous visual augmented feedback regarding the accuracy of speech movements is expected to help speakers to relearn the sound articulation and/or to learn adaptive strategies, resulting in improved speech.

Visual information is an essential component of speech processing by humans. A wealth of developmental literature suggests the importance of visual cues for infants acquiring speech [6]. The availability of a visual channel (i.e., face, head and even tongue movements) during communication significantly improves speech intelligibility [7,8]. Visual information has been successfully used in various speech treatments of voice and nasal resonance (see reviews in [9,10]). Movement-based techniques focused on the tongue, however, are not well developed or commonly used in speech therapy. Movement-based augmented feedback has been shown to be highly effective in neuromotor rehabilitation in the limb system [11,12]. Recent studies of limb motor control showed that visual information was important for improving the success rate of movements and decreasing variability near the end of the movement [13,14].

Considerable technological development has been required in order for augmented visual feedback to be implemented as a practical rehabilitation tool for the tongue. Current options for the sensing of lingual articulation include electropalatography (EPG, Articulate Instruments, UK) and electromagnetic articulography (2D EMA, Carstens Medzinelectronic, Germany). Both systems are well-developed and very sophisticated [15,16]; however, they have disadvantages for clinical use with patients with motor speech disorders. The EPG system requires the use of an artificial palate (typically an acrylic plate in which a sensor array has been embedded), which significantly alters sensory feedback during training. The palate needs to be custom-built for each patient, adding inconvenience and expense. The 2D EMA that has been used in the past requires the use of a heavy helmet; this is not ideal for clinical populations with generalized muscle weakness. A recently-developed 3D model does not make use of a heavy helmet, yet is still relatively difficult to use, as it requires extensive calibration and specialized training. A new technique for sensing lingual articulation has recently become available, however. The Wave system (Northern Digital Inc., Waterloo, Canada) is capable of tracking a patients tongue and other speech articulators in the oral cavity with sub-millimeter accuracy [4]. The Wave system has the advantage of being minimally invasive (e.g., its ultra-small 3mm sensors are mounted inside the oral cavity using dental adhesive) and user friendly. (See section 4.1 for further illustration.) This device has great potential to become a standard clinical tool for speech therapy.

3 Augmented Visual Feedback

Our system offers two layers of user feedback: (1) synchronous feedback that is provided during game play, implemented by the responsivity of the game elements to the actions of the game controller (the client's own tongue), and (2) "task completion" feedback, which conveys the degree to which the speech task was completed successfully. One of our system's distinguishing features is with respect to the first type of feedback. Systems used in other computer-aided speech domains, such as Computer Aided Language Learning (CALL) and Computer Aided Pronunciation Training (CAPT), do not typically have access to real-time kinematic data and thus must make use of acoustic-to-articulatory inferencing instead. This inferencing process entails reverse-engineering the articulation of the vocal tract (or at least the positions of the most salient articulators) from an acoustic source. The challenges of acoustic-to-articulatory inversion are well established, and many promising techniques have been developed (e.g., see [17]). However, the fact remains that, in present forms of computer-based feedback, the accuracy of information about the shape of the articulatory track is subject to noise. Moreover, this information is not available in real-time, since it is the product of computationally expensive algorithms. In turn, these factors (noise, latency) are unavoidably propagated into the corrective feedback that the system provides to the user.

Our hypothesis is that our system, which is based on real-time tongue motion data and employs mechanisms of computer game play, will have multiple benefits for patients. In particular, we hypothesize that the system will elicit a heightened and clinically-relevant level of *embodied interaction* — that is, the system will leverage the user's experience from a realm of abstraction and ideas to a realm of salient bodily experience [18]. *Salience of training* is a key principle in neural plasticity for rehabilitation in speech motor control [19]. In addition, we hypothesize that several factors (such as the immediacy of feedback and the intuitiveness of the gaming scenarios) will converge so as to elicit higher levels of motivation that are sustained longer than traditional modes of therapy. The possibilities that such a system offers give rise to very interesting research questions. In particular we are trying to set the foundation for answering important questions such as what is the right kind of feedback for speech therapy patients? Is a first person view more effective that a third person view? Is a competitive scenario more effective than a cooperative one?

4 Framework

We have designed a framework to enable the rapid development of interactive games for speech rehabilitation. Our framework is based on the integration of a game engine with the Wave motion tracking system. The Wave system is responsible for real-time tracking of the sensors attached on a patient's tongue. The collected measurements are streamed into our middle-layer, which processes the motion data and generates movement instructions for game objects. These instructions are then sent to the game engine, which renders the visualization on the patient's display.

The architecture of the framework, presented in Figure 1, consists of the following components:

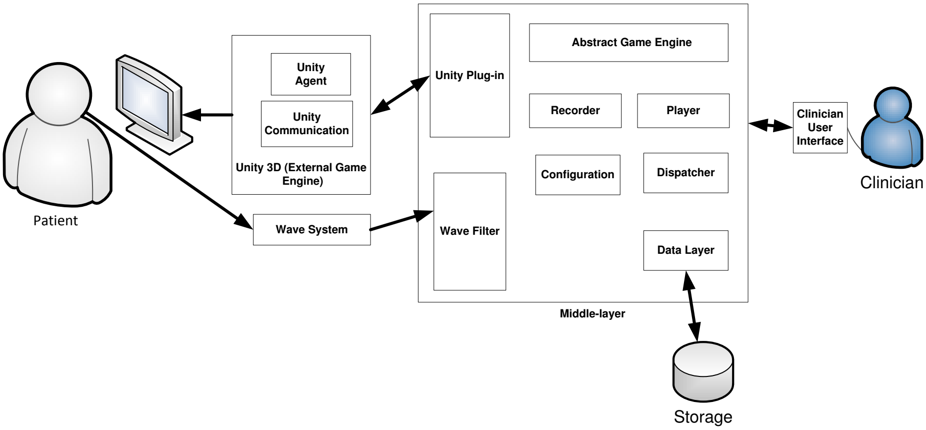

Fig. 1. Overview of our framework

- **Clinician User Interface (UI)**: A module that allows a clinician to configure a training session, select a particular game from game library, input patient information and set the parameters of the recording session. It also allows clinicians to review recorded sessions.
- **Middle Layer**:
  - **Unity Plug-in**: The game engine driver and the main interface between the middle-layer and the Unity game engine.
  - **Wave Filter**: A series of filters that transform and clean the raw motion data into usable control signals.
  - **Dispatcher**: A module that is responsible for distributing high level data between the middle-layer components and the external network client (i.e. the clinician).
  - **Data Layer**: A module that implements the main database operations for storing traces, recorded sessions, and patient–related information.
  - **Abstract Game Engine**: A layer that serves to make the framework game engine independent from the other middle-layer modules.
  - **Configuration**: A module that configures each session on the basis on the parameters set by the clinician.
  - **Recorder**: A module that records each training session and stores its relevant data in the main database.
  - **Player**: A module that affords replay of any previously recorded session that is stored in the main database.
- **Unity 3D (External Game Engine)**: The game engine (commercially-provided).
  - **Unity Communication**: An abstraction of the communication with our middle layer.
  - **Unity Agent**: A module that translates between the gaming instructions in the abstract game engine's format and the Unity3D scripts.
- **Wave System**: Sensor system (commercially-provided).

4.1 Motion Tracking

The Wave system [4] is an electromagnetic device designed to collect position of 5 and 6 degrees of freedom (DOF) sensors in a 30 × 30 × 30cm field. The average error obtained with Wave (including the head correction procedure) was reported at < 0.5 mm in the 300 mm^3 field volume [4]. Error of this size is acceptable in other articulo-graphic devices (see result of an assessment of the EMA system in [20]). When sensors are attached to the tongue, jaw, and lips, their movements can be tracked relative to the head. The Wave system includes a digital probe with a 6DOF sensor located at its tip. The probe can be used to trace any surface that is located within the recording field, e.g. the hard palate. Real-time visualization of the sensors allows on-line monitoring of the procedures, including positioning of the subject's head in the center of the tracking field. Wave collects movement data at $100, 200$ or 400 Hz and outputs positional (x, y, z) and orientation data (quaternion) at each time sample.

Work completed by our collaborative team has recently established the validity of using the probe to derive an accurate model of the palate [21]; this will be a key aspect for tongue trajectory rehabilitation that involves modelling contact with the palate (e.g., recovery of the ability to produce certain classes of consonants).

Figure 2 shows a patient with a single sensor attached to his tongue and his head next to the Wave system's tracking mechanism. This is the set up we use for our initial experiments.

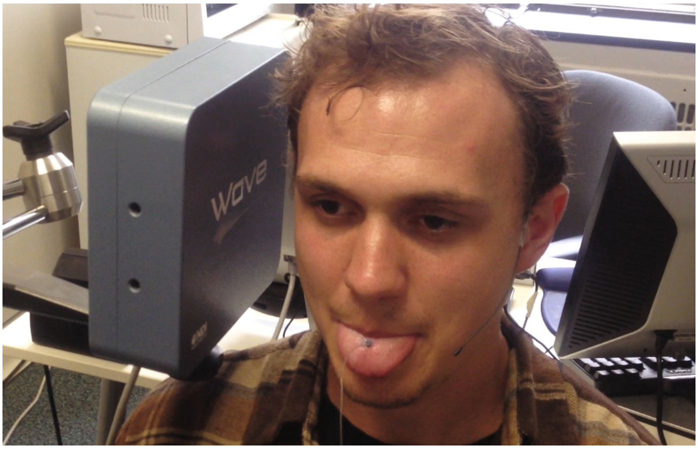

Fig. 2. The wave system and one of its sensors attached to a person's tongue

4.2 Unity Game Engine

Unity3D is a sophisticated tool for developing high quality interactive computer games [22]. It offers all the necessary components so that anyone with basic programming skills can quickly prototype complex interactive games. Among its most impressive features is its ability to automatically compile and deploy a game in different platforms, such as PCs, mobile phones and web-browsers. Unity offers a powerful scripting interface that supports Javascript, Boo and C# under the free license, as well as a plugin API in C++ exclusively available with the commercial license. Unity offers effective and clean interfaces through which developers can connect data streaming devices, game controllers and external processes, locally or through a network. Furthermore, there is a large development community around Unity that has created an impressive amount of content and assets that are publicly available. For these reasons, we have decided to build our current prototype framework based on Unity. However, there are other game engines that one could use for this purpose. The abstract game engine module (see Figure 1) allows our framework to use any game engine that is similar to Unity3D.

5 Usage and Initial Experiments

Our system is used as follows: the client is given a set of repetitive speech exercises, such as to produce a particular speech sound ten times. Each time the client produces the speech sound, his or her performance animates the game scenario. A correct speech production has an analog in the game-world that corresponds to the correct completion of a task in the game world. In effect, the client uses his or her speech apparatus as the game controller, and the game play is scaffolded by the speech language pathologist (SLP) implementing a clinical protocol for speech rehabilitation. The following sections present initial experiments towards creating simple gaming scenarios for such use.

5.1 Neutral Feedback

A proof-of-concept prototype was developed in which the user controlled the vertical location of a wooden plank using the motion of his tongue, Figure 3. In this particular game scenario, the wooden plank was placed on a deformable red ball. Lowering the tongue lifts the plank, whereas raising the tongue lowers the plank. The system currently provides two views: the first-person view that is achieved by positioning a virtual camera on the plank, and third-person view that is acheived with the simulated camera positioned above and to the right of the plank. For the sake of illustration, a close-up of the screen is shown in Figure 4. Unity3D provides the simulated dynamics along with a simple interface to create multiple cameras and split views.

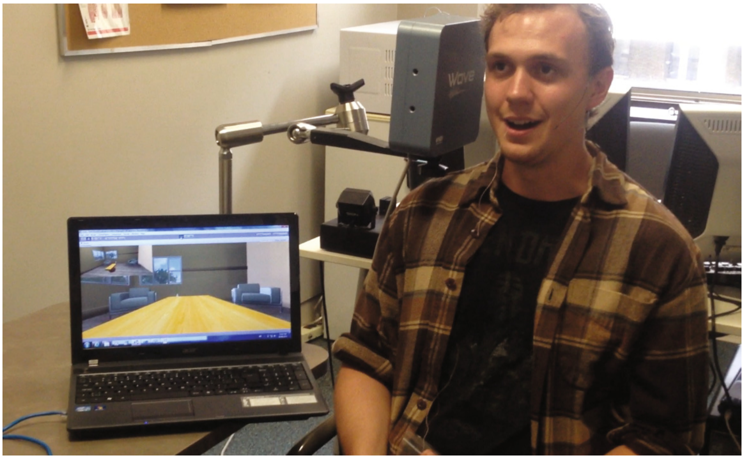

Fig. 3. Neutral feedback: A simulated wooden plank, resting on a flexible red ball, rises and lowers following the movement of a patient's tongue. The main image on the screen shows the first-person and third-person perspective simultaneously.

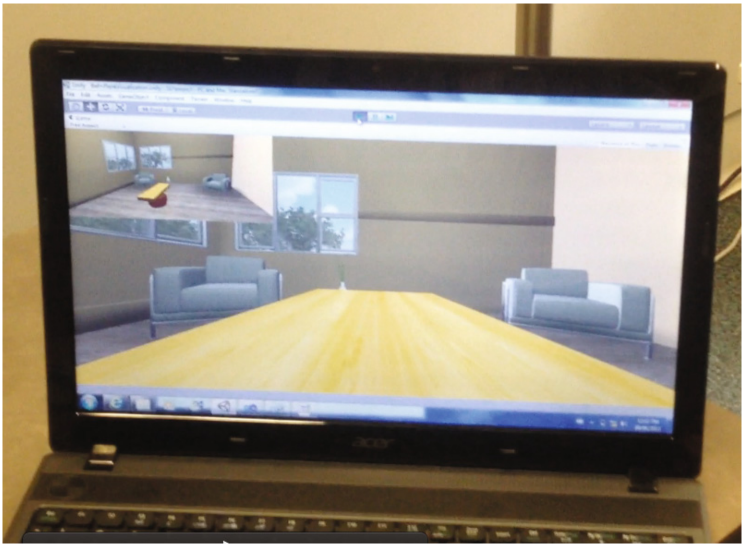

Fig. 4. A patient's views: first-person view in fullscreen, and third-person view in the inset in the upper left corner

5.2 Success/Failure Feedback

In the subsequent iteration of our system, the main game scenario provides a cartoon-looking bee avatar, which the game player controls in order to land on flowers and collect pollen. Figure 5 shows snapshots of the prototype environment that we developed for this experiment.

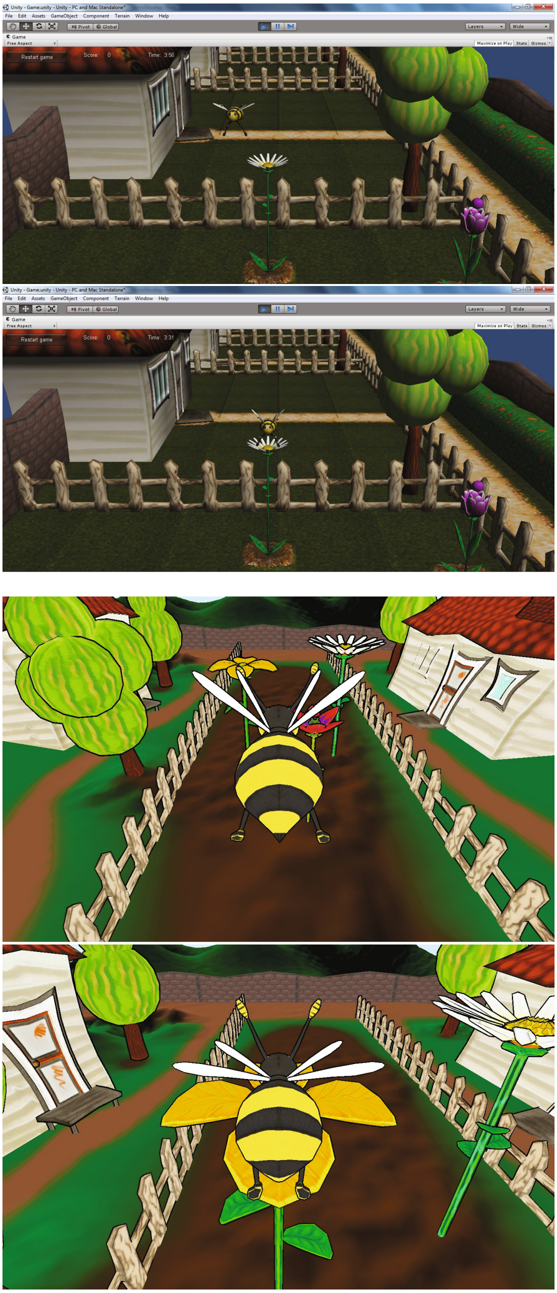
Unity - Game.unity - Unity - PC and Mac Standalone*
File Edit Assets GameObject Component Terrain Window Help
Pivot Global
Layers Wide
Game
Free Aspect
Maximize on Play Stats Gizmos
Restart game
Score: 0
Time: 3:56
Unity - Game.unity - Unity - PC and Mac Standalone*
File Edit Assets GameObject Component Terrain Window Help
Pivot Global
Layers Wide
Game
Free Aspect
Maximize on Play Stats Gizmos
Restart game
Score: 0
Time: 3:31

Fig. 5. Snapshots from the bee game showing a two-frame sequence of a user successfully guiding the bee onto the flower using his or her tongue (in each of two perspectives). The frames on the top illustrate the third-person view; the frames on the bottom illustrate an over-the-shoulder first-person view with cartoon shading.

In this version, a patient uses his or her tongue to guide the bee along the cardinal directions. Basic tongue motion is an element of the early phases of rehabilitation post-stroke. Each successful task completion earns the user points. The clinician can tailor the game to the specific abilities and needs of each patient. For example, the game can be simplified to allow only lateral movement (the patient just moves the bee up and down, or left and right), or it can be enhanced with the addition of obstacles (around which the bee must manoeuvre) and/or "enemy" characters, such as spiders (which can be autonomous or clinician-controlled). In the experiment shown in Figure 5, the user has successfully moved the bee from its initial position down onto the flower. Two sets of frames are shown, in each of the first-person and third-person views, which showcases the ability of the game engine to easily customize the look and feel of a game.

6 Discussion

We have presented work in progress towards developing a computer gaming framework for speech rehabilitation. To our knowledge, this is the first attempt to bring together both real-time kinematic tracking of the speech articulators and the well-established power and appeal of gaming to the realm of speech rehabilitation. This combination of state-of-the-art technologies opens up many exciting possibilities, which must be explored pragmatically, as there remain many important open research questions about the characteristics of effective feedback mechanisms and about eliciting salient experience through user interaction. Our immediate next steps are to investigate the following issues:

- Impact of view: To what degree does view (i.e., first-person vs third-person) impact the elicitation of experiences that are effective for rehabilitation? If first-person view brings about stronger treatment effects (as we hypothesize), is the embodiment in interaction the explanatory mechanism?
- Role and nature of competition: Individual task scenarios will require patients to complete a goal-oriented task using their tongue. The elicitation of target tongue trajectories will be presented visually, using tasks such as asking the user to correctly position a visual element or hitting a particular target within the game scenario. Competitive scenarios can implement progressive levels of challenge for patients as they improve and can be implemented via gaming AI approaches (with input from clinicians). Which aspects of competition are most salient to avoiding demotivation? Which aspects of competition are most salient for invoking and maintaining the targeted speech behaviour (correct articulation)? To what extent, if any, do the goals of demotivation avoidance and behaviour modification conflict?

To answer these questions, we plan to employ both qualitative and quantitative user study methodologies in a population of adult stroke patients and their clinicians. Our long-term goal is to evaluate the relative efficacy of the different views and different competition modes using a variety of measures, including behavioural outcomes (i.e., the progress of patients toward clinical goals) and subjective measures (e.g., through questionnaires and interviews with the patients and clinicians).

Acknowledgements. The authors gratefully acknowledge funding support from the NSERC Discovery Grant Program, the ASHA Foundation (New Investigator Award), and the Centre for Innovation in Information Visualization and Data-Driven Design (CIV-DDD).

References

1. Engwall, O., Bälter, O.: Pronunciation feedback from real and virtual language teachers. Computer Assisted Language Learner 3, 235–262 (2007)
2. Vicsi, K., Roach, P., Oster, A., Kacic, Z., Barczikay, A., Tantoa, A., Csatari, F., Bakcsi, Z., Sfakianaki, A.: A multilingual teaching and training system for children with speech disorders. International Journal of Speech Technology 3, 289–300 (2000)
3. Piper, A., O'Brien, E., Morris, M., Winograd, T.: Sides: a cooperative tabletop computer game for social skills development. In: Proceedings of the 2006 20th Anniversary Conference on Computer Supported Cooperative Work, CSCW 2006, pp. 1–10. ACM, New York (2006)
4. Berry, J.: Accuracy of the NDI wave speech research system. Speech Language, and Hearing Research 54, 1295–1301 (2011)
5. Massaro, D., Bigler, S., Chen, T., Perlman, M., Ouni, S.: Pronunciation training: the role of eye and ear. In: Proceedings of Interspeech 2008, pp. 2623–2626 (2008)
6. Kuhl, P.K., Meltzoff, A.N.: The bimodal perception of speech in infancy. Science 218, 1138–1141 (1982)
7. Lucero, J., Maciel, S., Johns, D., Munhall, K.: Empirical modeling of human face kinematics during speech using motion clustering. The Journal of the Acoustical Society of America (January 2005)
8. Mcgurck, H., Macdonald, J.W.: Hearing lips and seeing voices. Nature 264(246-248) (1976)
9. Davis, S.M., Drichta, C.E.: Biofeedback theory and application in allied health: speech pathology. Biofeedback Self Regul. 5(2), 159–174 (1980)
10. Maryn, Y., De Bodt, M., Van Cauwenberge, P.: Effects of biofeedback in phonatory disorders and phonatory performance: a systematic literature review. Appl. Psychophysiol. Biofeedback 31(1), 65–83 (2006)
11. Mandel, A., Nymark, J., Balmer, S., Grinnell, D., O'Riain, M.: Electromyographic versus rhythmic positional biofeedback in computerized gait retraining with stroke patients. Arch. Phys. Med. Rehabil. 71(9), 649–654 (1990)
12. Moreland, J., Thomson, M.A.: Efficacy of electromyographic biofeedback compared with conventional physical therapy for upper-extremity function in patients following stroke: a research overview and meta-analysis. Phys. Ther. 74(6), 534–543 (1994); discussion 544–547
13. Beacutedard, P., Proteau, L.: On-line vs. off-line utilization of peripheral visual afferent information to ensure spatial accuracy of goal-directed movements. Exp. Brain Res. 158(1), 75–85 (2004)
14. Franklin, D.W., So, U., Burdet, E., Kawato, M.: Visual feedback is not necessary for the learning of novel dynamics. PLoS ONE 2(12), e1336 (2007)

15. McNeil, M., Katz, W., Fossett, T., Garst, D., Szuminsky, N., Carter, G., Lim, K.: Effects of on-line augmented kinematic and perceptual feedback on treatment of speech movement in apraxia of speech. Folia Phoniatr. Logop. 62, 127–133 (2010)
16. Hardcastle, W., Gibbon, F., Jones, W.: Visual display of tongue palate contact: electropalatography in the assessment and remediation of speech disorders. Br. J. Disord. Commun. 26, 41–74 (1991)
17. Engwall, O.: Analysis of and feedback on phonetic features in pronunciation training with a virtual teacher. Computer Assisted Language Learning 25(1), 37–64 (2012), QC 20120416
18. Dourish, P.: Where the action is: the foundations of embodied interaction. MIT Press, Cambridge (2001)
19. Ludlow, C.L., Hoit, J., Kent, R., Ramig, L.O., Shrivastav, R., Strand, E., Yorkston, K., Sapienza, C.M.: Translating principles of neural plasticity into research on speech motor control recovery and rehabilitation. J. Speech Lang. Hear. Res. 51(1), 240–258 (2008)
20. Yunusova, Y., Green, J., Mefferd, A.: Accuracy assessment for AG500, electromagnetic articulograph. Speech Language, and Hearing Research 52, 547–555 (2009)
21. Yunusova, Y., Baljko, M., Pintilie, G., Rudy, K., Faloutsos, P., Daskalogiannakis, J.: Acquisition of the 3d surface of the palate by in-vivo digitization. Speech Communication 54, 923–931 (2012)
22. Unity Inc.: Unity3D, http://unity3d.com/unity/