INTERSPEECH 2016
September 8–12, 2016, San Francisco, USA

Sensorimotor response to visual imagery of tongue displacement

William F. Katz [1], *Divya Prabhakaran* [2]

[1] The University of Texas at Dallas
[2] Plano East Senior High School

wkatz@utdallas.edu, divya.prabhakaran6@gmail.com

Abstract

To better understand audiovisual speech processing, we investigated the effects of viewing time-synchronized videos of a 3D tongue avatar on vowel production by healthy individuals. A group of 15 American English-speaking subjects heard pink noise over headphones and produced the word *head* under four viewing conditions: First, while viewing repetitions of the same vowel, /ɛ/ (baseline phase), then during a series of "morphed" videos shifting gradually from /ɛ/ to /æ/ (ramp phase), followed by repetitions of /æ/ (maximum hold phase), and finally repetitions of /ɛ/ (after effects phase). Results of a formant frequency (F1) analysis indicated that the visual mismatch phases (ramp and maximum hold) caused all subjects to align their productions to the visually-presented vowel, /æ/. No subjects reported being aware that their vowel quality had changed. We conclude that the visual moving tongue stimuli produced entrainment to the viewed vowel category, rather than adaptation in the opposite direction of the perturbation. Further experimentation is needed to determine whether these effects are due to inherent imitation behaviors or subjects' lack of agency with the tongue avatar.

Index Terms: speech production and perception, visual feedback, electromagnetic articulography, sensorimotor adaptation

1. Introduction

Speech communication involves the sensorimotor integration of auditory, tactile, orosensory, and visual information [1, 2]. An important means of investigating sensorimotor integration in speech is to conduct feedback perturbation experiments, in which sensory information is altered so that underlying control processes and short-term learning may be observed [3, 4, 5]. Perturbation delivered in an unexpected and random fashion is assumed to tap moment-to-moment control processes (compensation), while perturbation applied in a more predictable and constant manner is thought to assess a form of short-term learning (adaptation).

Several acoustic feedback studies have recently investigated vowel production by having subjects hear their voice (mixed with noise) over headphones while a rapid, online acoustic perturbation that changes the status of one or more speech parameters is introduced (fundamental frequency [F_0], formant frequencies, or amplitude). Sensorimotor *compensation* experiments have generally found that subjects can rapidly adjust in the opposite direction from the perturbation. This has been noted for shifts in formant frequencies [3,4] and F_0 [6,7]. Similarly, sensorimotor integration serves as the basis for procedural learning, involving adaptive motor changes for altered sensory cues [8, 9, 10, 11]. Like the findings for compensation, sensorimotor *adapation* experiments have demonstrated changes in the opposite direction to the feedback shift [5, 11]. Also, in these adaptation experiments, when normal feedback is suddenly restored there is typically a shift towards feed-forward (rather than feed-back) planning.

Taken together, acoustic feedback perturbation studies suggest that vowel goals are primarily auditory in nature and that both immediate control processes as well as short-term learning act together to maintain vowel phonetic quality during speech. It is important to note that these studies have been restricted to the effects of auditory feedback; that is, on-line shifting of either the F_0, formant frequencies, or amplitudes of speech signals delivered acoustically to subjects during speech. However, it is well-known that speech frequently involves both the auditory and visual modalities [1, 2, 12]. It therefore remains unclear whether other sensory modalities (e.g., vision) play a role in speech compensation or adaptation.

Recent studies using "mirror" and "silent mouthing" conditions during speech have supported the view that combinations of visual and auditory speaking conditions can affect speech perception and/or production. For instance, watching one's own face can induce McGurk effect-type blends and simultaneous silent articulation of a concordant stimulus moderately improves auditory comprehension [13]. Similarly, silently articulating a syllable in synchrony with the presentation of a concordant auditory and/or visually ambiguous speech stimulus appears to improve syllable identification, with concurrent mouthing further speeding the perceptual processing of a concordant stimulus [14, 15, 16]. Overall, these studies indicate that listeners benefit from multimodal speech information (including knowledge from one's own motor experience) during the perception process.

In terms of specific findings for audiovisual compensation or adaptation, little is known. As a first step in addressing this question, we investigated the effect of having subjects produce a vowel while viewing an avatar representing the movement of their tongue. We chose the tongue because "tongue reading" studies using avatar-based instructional systems, such as Baldi [17, 18] or ARTUR [19, 20], have shown small but consistent perceptual improvement when tongue movement information is added to the visual display. Positive effects have been noted in word retrieval for acoustically degraded sentences [21] and in a forced-choice consonant identification task [22]. Also, recent findings from systems providing interactive tongue movement information to subjects during speech have reported benefits in novel speech sound training [23, 24, 25, 26].

In the present research, a visual feedback perturbation (*adaptation*) experiment was conducted in which the vowel /ɛ /

http://dx.doi.org/10.21437/Interspeech.2016-1594

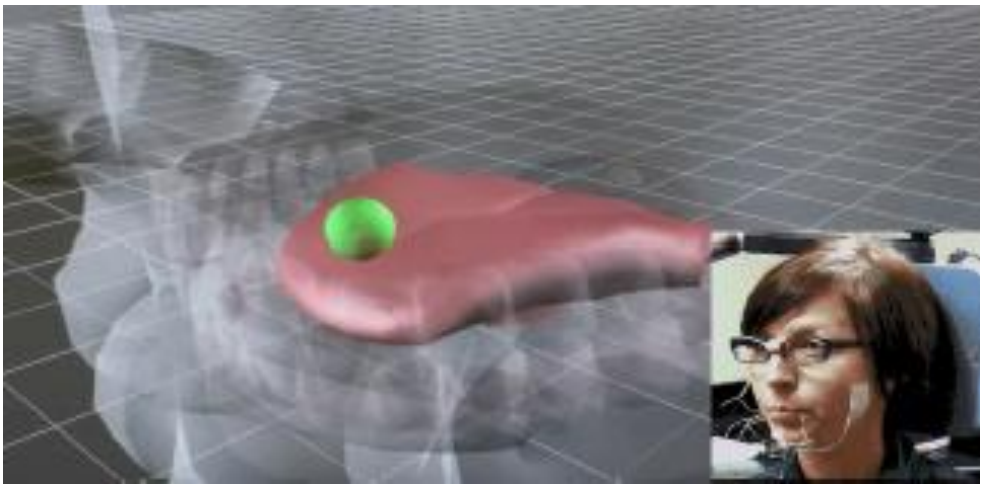

Figure 1. *Screenshot from the Opti-Speech system, with a subject wearing sensors and head-orientation glasses*

was gradually morphed to the more open vowel /æ/ (ramp phase), then returned back to /ɛ/ again. Based on previous findings suggesting that listeners benefit from multimodal speech information during the perception process [16] and that audiovisual speech can also influence production [27], we predicted that the visual tongue imagery would influence talkers' vowel productions. Specifically, we predicted that talkers would show adaptation in the opposite direction of the visual shift (i.e., toward increased /ɛ/ vowel quality). Also, with return to visual /ɛ/, it was expected that subjects would adopt more feed-forward processing and rapidly return to previous (pre-shifted) values.

2. Methods

2.1. Participants

Fifteen speakers (12 female) between the ages of 18 and 26 years old, volunteered to participate in the experiment. All were monolingual speakers of American English from the University of Texas at Dallas community. None reported any history of speech, hearing, or language disorders. None had any experience with the virtual tongue model.

2.2. Visual Stimuli

The experiment used images from an animated 3D tongue avatar, with data captured from actual tongue movements produced by a male native speaker of American English (WK) speaking the words *hid, head,* and *had*. Tongue movements were visualized using an interactive articulatory feedback system, Opti-Speech [27], based on data recorded using the WAVE magnetometer system (Wave; NDI, Waterloo, Ontario, Canada). The Opti-Speech system allows a speaker to view his/her current tongue position (represented as an avatar consisting of flesh-point markers and a modeled surface) placed in a synchronously moving, transparent head (Figure 1). The contrasting vowel stimuli (/ɛ/,/æ/) were chosen because they correspond with easily observed tongue movements and they have yielded robust shifts in previous perturbation experiments. Video editing software (Camtasia 2, Techsmith, 2015) was used to record moving images of the tongue model while the /hVd/ words were produced. Animation software (Adobe After Effects, Adobe Systems, 2015) was subsequently used to morph video clips of the tongue avatar in a five-step continuum from *head* to *had*. In order to encourage simultaneous speech production while viewing the avatar tongue movements, each /hVd/ video clip was preceded by a "3,2,1" countdown and a green "get ready" signal (Figure 2). Readers can find a movie demonstrating the five-stage morphed video continuum (from *head* to *had*) along with a PPT file containing a sample of the warmup trialsat http://www.utdallas.edu/~wkatz/research.html.

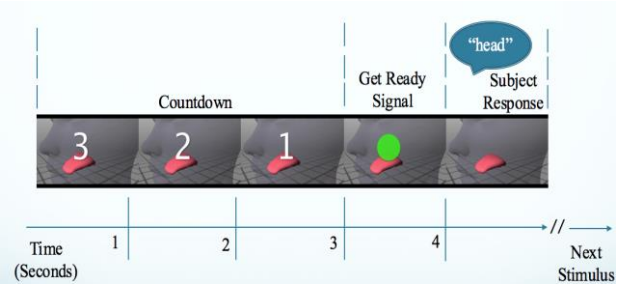

Figure 2. *Overview of the synchronized tongue viewing and speaking task.*

The video materials were assembled in timed presentations for playout (both for a warmup trial and the actual experiment). During the experiment, stimuli were shown in blocks of 10, with a one second inter-stimulus interval (ISI) between video clips, and a five second inter-block interval (IBI). The entire speaking task took approximately 17 minutes.

2.3. Procedure

Each participant was seated facing a computer monitor while wearing closed-cell headphones (Sennheiser HD 500) which transmitted masking noise at approximately 72 dB. Pink noise (i.e., having spectral power density decreasing by 3 dB per octave) was selected based on user comments indicating this type of masking noise was comfortable to listen to during the

main experiment. Participants were first given instructions regarding the experimental procedure and introduced to the task through a series of warmup trials. Next, the experimental trials were begun, consisting of four phases in sequential order (baseline/ramp/maximum hold/after effects). The recordings of productions of *hid, head,* and *had* were elicited for baseline and vowel normalization purposes. Five repetitions were elicited for each vowel at baseline. In this condition, talkers produced movements in concordance with each spoken word. During the ramp phase, participants were asked to say *head* in time with the morphed video images ranging from *head* to *had* (five stages, eight repetitions each). This instruction was provided in writing on the computer monitor ("The next words will be *head*") before the ramp phase. In a similar fashion, during the maximum hold phase, participants were asked to say *head* while watching the tongue avatar movement for *had* (five sets, 20 repetitions each). In the after effects phase, participants were also asked to say *head* in synchrony with *head* visual images, (two stages, 15 repetitions each). A Tascam DR-05 recorder was used to record audio data.

After the speaking experiment finished, participants were debriefed by being asked "What did you notice about this experiment?" The purpose of this question was to obtain participants' impressions concerning the difficulty of the task and to discern whether participants were aware that the visual tongue positions had changed vowel quality. After recording participants' initial responses, we next informed participants that the avatar had actually shifted from /ɛ/ to /æ/, and participants were further queried whether they were aware of such a change taking place.

2.4. Acoustical analyses

For the 15 total productions of the baseline words (*hid, head, had*) and 185 productions of the target word *head,* linear predictive coding (LPC) in Praat [28] was used to estimate first formant (F1) frequencies at the vowel midpoint. The data were normalized using Lobanov's *z*-score transformations to reduce variation due to male/female vocal tract anatomical differences [29]. Averaged *F*1 values were then compared across the four different phases (baseline/ramp/maximum hold/after effects) of the experiment.

3. Results

The averaged F1 /ɛ/ vowel productions for the n=12 female talkers are shown for the four test phases in Figure 3. The expected F1 for /ɛ/ (based on Texas female talkers [30]) is indicated by the lower dotted line, and /æ/ by the top dotted line. The data suggest that talkers first produced rather typical /ɛ/ F1 values (mean of 734 Hz at baseline), then approached /æ/ during the ramp phase (peak mean of 755 Hz; a 9.4% increase). Formant frequency values remained high during the maximum hold phase (peak mean of 765 Hz; a 10.8% increase over baseline), then lowered again during the after effects phase, although not quite returning to the original /ɛ/ level.

The F1 data for the three male talkers are shown in Figure 4. Due to the low number of participants, standard errors are not shown in this Figure. The expected F1value for /ɛ/ (based on Texas male speakers [30]) is also indicated. Results suggest a similar pattern to the females: talkers produced typical /ɛ/ *F*1 values (mean of 616 Hz) at baseline, then approached /æ/ during the ramp phase (peak mean of 632 Hz; a 2.4% increase). Formant frequency values continued to increase in the maximum hold phase (peak mean of 671 Hz; a 8.9% increase over baseline), then lowered markedly during the after effects phase, a 4.9% decrease from baseline.

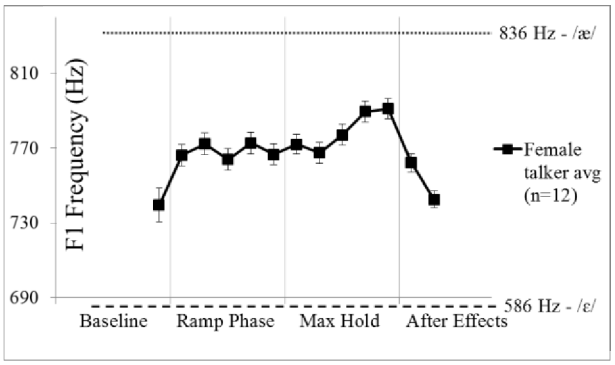

Figure 3: *Average* F1 *frequency of n=12 female talkers across the four speaking conditions*. *Error bars show standard errors.*

The normalized data for the *n*=15 talkers were tested statistically in a one-way, repeated measures analysis of variance (ANOVA) comparing the effects of experimental condition (baseline/ramp/hold/after effects) across talkers. The results indicated a significant main effect for condition [$F(3, 42) = 6.94$, $p < 0.001$], with *Bonferronni*-adjusted contrasts indicating significant differences between baseline and maximum hold ($p < 0.05$), ramp and maximum hold ($p < 0.005$), maximum hold and after-effects ($p < 0.001$), and ramp and after-effects ($p < 0.05$).

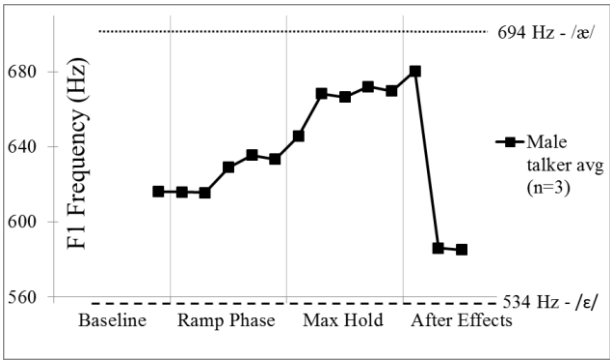

Figure 4. *Average* F1 *frequency of n=3 male talkers across the four speaking conditions. Error bars show standard errors.*

Upon debriefing, no participants indicated they were aware that the tongue avatar had switched from /ɛ/ to /æ/. Some participants mentioned that the /hVd/ words "ended differently", others thought there were "trick sounds" being played in the noise, while a few commented on the fact that the tongue avatar position visibly changed "once or twice." When the participants were informed that the avatar had actually shifted to an /æ/ they were also asked whether they were aware of having produced this vowel: All replied "no" – participants reported being only aware of producing the vowels /ɪ/ and /æ/

at the beginning of the experiment (baseline phase), then /ɛ/ thereafter.

4. Discussion

In order to examine how visual information influences vowel processing during speech, a group of 15 talkers participated in an on-line adaptation paradigm in which words were produced while viewing a concurrent, moving tongue image. It was predicted that the tongue image would influence the vowel quality of the talkers' productions (presumably the result of also influencing their perceptions) and that this would cause a shift away from the perturbing stimulus. Thus, as the /ɛ/ stimulus shifted towards /ae/, participants would adapt, further raising the jaw towards /ɛ/ /and thereby causing a lowered F1 frequency.

Our first prediction was only partially met, in that participants' F1 values changed as a function of the experimental conditions. However, listeners did not report hearing a change in vowel quality as the result of viewing the tongue, and the direction of the spoken productions was unexpected: Values shifted *towards* the perturbing stimuli. That is, during the ramp phase, talkers' /ɛ/ became more /æ/-like, were maintained at /æ/-like values during maximum hold, and then returned to /ɛ/ values during the after effects phase.

Concerning the second prediction, there was a statistically significant return to /ɛ/ vowel formant frequency values from the maximum hold to the after effect phases for the talkers. This effect appeared to be stronger for the men than the women, although this sex difference was not tested statistically. Taken together, the data suggest that rather than adapting, all subjects appeared to entrain to the visual avatar.

Before considering possible explanations, one concern might be accounting for speech variation caused by the many potential factors involved in the experimental setup (including the participants' need to speak in masking noise and to follow the tongue avatar). That is, perhaps the participants produced highly unnatural speech, raising issues of validity. To investigate this possibility, we obtained a total of 10 spontaneous speech samples of *hid*, *head*, and *had* from a different group of adult participants before, during, and after a very similar tongue morphing experiment. These participants' F1 values were compared with samples obtained from their baseline (concordant, *head*) and maximum shift (discordant, *had*) experimental productions. The overall difference between these talkers' spontaneous speech samples and their vowel experimental data was 2.3% (0.0-9.0%). This close match in F1 values between spontaneous and experimental samples suggests that talkers are not artificially constrained during the speaking task.

There are at least two (non-competing) explanations for the current findings. First, several lines of research support a strong biological basis for the mirroring or "mimicking" of tongue movement behavior. Tongue protrusion is a widely-studied imitative gesture, found to produce statistically significant effects in infants as young as 2-3 days old [31]. A large (and rather controversial) literature also describes a "mirror neuron" system for humans, based on premotor and parietal cells in the primate brain that fire during the performance of an action and when seeing others perform that same action [32, 33]. Accordingly, MEG studies have found links between human brain regions controlling tongue motor and speech perceptual areas [34]. In addition, a series of intervention studies have reported positive findings for the use of audiovisual (facial) imitation as a means of remediating the expressive speech disorders of Broca's aphasia [35, 36]. Taken together with the present findings, these data suggest that visual information relating to tongue movement results in behavior that is strongly imitative and therefore plays no role in adaptation. Thus, the participants in the present experiment watched the tongue avatar and mimicked it. This behavior changed their vowel quality, although due to noise masking, they were unaware that this qualitative change had taken place.

A second possible explanation for the current finding of entrainment (rather than adaptation) may be that participants were not sufficiently convinced a perturbation was in fact taking place. That is, due to the methodological constraint of requiring a morphed tongue image to be displayed on a monitor, participants were aware that the tongue avatar movements were not their own. In this way, the present experiment diverged from previous acoustic perturbation paradigms in which the talker's own acoustic values are altered in real-time, and the listener has every reason to suspect the altered stimulus is his/her own. In future studies, we propose to conduct on-line tongue perturbations using the Opti-Speech avatar, to test the possibility that an increased sense of agency may be necessary for putative adaptation effects to be observed.

5. Conclusions

The present findings indicate that talkers continue to hear *head* but say *had* when presented with a gradually changing image of a tongue morphing from *head* to *had* (and while tasked with saying *head*). We conclude that in this case the visual stimuli led to entrainment, not adaptation. Future studies will be needed to determine how audiovisual information is integrated during speech motor control and short-term adaptation. Studies using on-line visual perturbations with articulatory feedback systems such as Opti-Speech may be particularly useful in this regard. These studies might also include stimuli more categorically perceived than vowels, such as consonants.

6. Acknowledgements

The research was completed as part of an Intel Science Fair Project by the second author. We thank the Plano East Senior High School for giving the second author release time for this project. We also thank Dr. Thomas Campbell and members of the University of Texas at Dallas C-Tech Center for providing resources for this research.

7. References

[1] W. H. Sumby and I. Pollack, "Visual contribution to speech intelligibility in noise," *J Acoust Soc Am,* vol. 26, pp. 212–215, 1954.

[2] B. Gick and D. Derrick, "Aero-tactile integration in speech perception," *Nature*, vol. 462, pp. 502–504, 2009.

[3] D. W. Purcell and K. G. Munchall, "Compensation folloing real-time manipulation of formants in isolated vowels," *Jasa*, vol. 119, pp. 2288, 2006.

[4] S, Cai, S. S. Ghosh, F. H. Guenther, and J. S. Perkell, "Focal manipulations of formant trajectories reveal a role of auditory feedback in the online control of both within-syllable and between-syllable speech timing," *J Neurosci*, vol. 31, no. 45, pp. 16483-16490, 2011.

[5] M. Shum, D. M. Shiller, S. R. Baum, and V. L. Gracco, "Sensorimotor integration forspeech motor learning involves the inferior parietal cortex," *Europ J Neurosci*, vol. 34, no. 11, 2011.

[6] J. J. Bauer and C. R. Larson, "Audio-vocal responses to repetitive pitch-shift stimulation during a sustained vocalization: Improvements in methodology for the pitch-shifting technique," *J Acoust Soc Amer*, vol. 114, no. 2, pp. 1048-1054, 2003.

[7] H. Liu and C. R. Larson, "Effects of perturbation magnitude and voice F0 level on the pitch-shift reflex," *J Acoust Soc Amer*, vol. 122, pp. 3671, 2007.

[8] J. F. Houde and M. I. Jordan, "Sensorimotor adaptation of speech I: Compensation and adaptation," *J Speech Lang Hear Res*, vol. 45, pp. 295-310, 2002.

[9] S. Tremblay, D. M. Shiller, and D. J. Ostry, "Somatosensory basis of speech production," *Nature*, vol. 423, pp. 866–869, 2003.

[10] J. A. Jones and K. G. Munhall, "Remapping auditory-motor representations in voice production," *Curr Biol*, vol. 15, no. 19, pp. 1768-1772, 2005.

[11] F. Mollaei, D. M. Shiller, and V. L. Gracco, "Sensorimotor adaptation of speech in Parkinson's disease," *Movement Disord*, vol. 28, pp. 1668-1674, 2013.

[12] L. Menard, "Multimodal speech production," in *The Handbook of Speech Production*, M. A. Redford. UK: J. Wiley, 200-221, 2015.

[13] M. Sams, R. Möttönen, and T. Sihvonen, "Seeing and hearing others and oneself talk," *Brain Research, Cognitive Brain Research,* vol. 23, pp. 429-35.

[14] M. Sato, E. Troille, L. ,Ménard, F. ,M.A. Cathiard,M.A.,and V. Gracco, V. "Silent articulation modulates auditory and audiovisual speech perception. *Exp.Brain Res*., vol. 227, pp. 275–288, 2013.

[15] T. Mochida,, T. Kimura, ,S. Hiroya, N. Kitagawa, H. Gomi,.,and T. Kondo. "Speech misperception: speaking and seeing interfere differently with hearing". *PLoSONE,* 2013.

[16] A. D'Ausilio, E. Bartoli, L. Maffongelli, J. J. Berry, and L. Fadiga, "Vision of tongue movements bias auditory speech perception," *Neuropsychologia*, vol. 63, pp. 85-91, 2014.

[17] D.W Massaro and M.M. Cohen, "Visible speech and its potential value for speech training for hearing-impaired perceivers," in STiLL-Speech Technology in Language Learning (Marholmen),1998.

[18] D. W. Massaro, "A computer-animated tutor for spoken and written language learning," in *Proceedings of the 5th International Conference on Multimodal Interfaces, pp.* 172-175*,* 2003.

[19] O. Engwall and O. Bälter, "Pronunciation feedback from real and virtual language teachers," *Comput. Assist. Lang. Learn,* vol. 20, pp. 235–262, 2007.

[20] O. Engwall, "Can audio-visual instructions help learners improve their articulation? – an ultrasound study of short term changes," in *Interspeech,* Brisbane, AUS, pp. 2631–2634, 2008.

[21] P. Wik and O. Engwall, "Can visualization of internal articulators support speech perception?" in *Proceedings of Interspeech,* Brisbane, AUS, pp. 2627–2630, 2008.

[22] P. Badin, Y.Tarabalka, F. Elisei, and G. Bailly, "Can you 'read' tongue movements? Evaluation of the contribution of tongue display to speech understanding", *Speech Commun. vol. 52*,493–503, 2010.

[23] J. L. Preston, P. McCabe, A. Rivera-Campos, J. L. Whittle, E. Landry, and E. Maas, "Ultrasound visual feedback treatment and practice variability for Residual speech sound errors," *J. Speech Lang. Hear. Res.*, vol. 57, pp. 2102–2115, 2014.

[24] W. F. Katz and M. R. McNeil, "Studies of articulatory feedback treatment for apraxia of speech based on electromagnetic articulography," *SIG 2 Perspect. Neurophysiol .Neurogenic Speech Lang. Disord*., vol. 20, pp. 73–79, 2010.

[25] W. F. Katz and S. Mehta, "Visual feedback of tongue movement for novel speech sound learning," *Frontiers in Human Neuroscience*, vol. 9, article 612, 2015.

[26] A. Suemitsu, T. Ito, and M. Tiede, "An electromagnetic articulography-based articulatory feedback approach to facilitate second language speech production learning," *Proceedings of Meetings on Acoustics – Acoustical Society of America*, vol. 19, no. 1, 2013.

[27] W. Katz, T.F. Campbell, J. Wang, E. Farrar, J. C. Eubanks, A. Balasubramanian, B. Prabhakaran and R. Rennaker. "Opti-speech: a real-time, 3d visual feedback system for speech training." *INTERSPEECH* (2014).

[28] P. Boersma and D. Weenink, (2010). Praat: doing phonetics by computer (Version 5.1.3) [Computer program]. Retrieved from [http://www.praat.org/].

[29] B. M. Lobanov,, "Classification of Russian vowels spoken by different speakers,'' *J. Acoust. Soc. Am.* 49, 606–608, ~1971.

[30] W. Katz and P. Assmann, "Identification of children's and adults' vowels: Intrinsic fundamental frequency, fundamental frequency dynamics, and presence of voicing," *Journal of Phonetics*, vol. 29, pp. 23-51, 2001.

[31] M. Heimann, K. E. Nelson, and J. Schaller, "Neonatal imitation of tongue protrusion and mouth opening: methodological aspects and evidence of early individual differences," *Scand. J. Psychol*, vol. 30, pp. 90–101, 1989.

[32] J. M. Kilner et al., "What We Know Currently about Mirror Neurons," *Current Biology*, vol. 23, no. 23, pp. R1057 - R1062, 2013.

[33] G. Hickok, "Eight Problems for the Mirror Neuron Theory of Action Understanding in Monkeys and Humans," *Journal of cognitive neuroscience*, vol. 21, no. 7, pp. 1229-1243, 2009.

[34] M. Sato, G. Buccino, M. Gentilucci, and L. Cattaneo. "On the tip of the tongue: Modulation of the primary motor cortex during audiovisual speech perception." Speech Communication 52, no. 6 pp. 533-541, 2010.

[35] J. Lee, R. Fowler, D. Rodney, L.Cherney, and S.L. Small, "IMITATE: An intensive computer-based treatment for aphasia based on action observation and imitation." *Aphasiology, vol.* 24, no. 4, 449-465, 2010.

[36] J. Fridriksson J.M. Baker, J. Whiteside, D. Eoute, D. Moser, R. Vesselinov, and C. Rorden, "Treating visual speech perception to improve speech production in nonfluent aphasia," *Stroke*. Mar 1;40(3), pp. 853-8, 2009. 2009.