Read my points: Effect of animation type when speech-reading from EMA data

Kristy James
University of Groningen
Saarland University
kristyj@coli.uni-saarland.de

Martijn Wieling
University of Groningen
The Netherlands
m.b.wieling@rug.nl

Abstract

Three popular vocal-tract animation paradigms were tested for intelligibility when displaying videos of pre-recorded Electromagnetic Articulography (EMA) data in an online experiment. EMA tracks the position of sensors attached to the tongue. The conditions were dots with tails (where only the coil location is presented), 2D animation (where the dots are connected to form 2D representations of the lips, tongue surface and chin), and a 3D model with coil locations driving facial and tongue rigs. The 2D animation (recorded in VisArtico) showed the highest identification of the prompts.

1 Introduction

Electromagnetic Articulography (EMA) is a popular vocal-tract motion capture technique used increasingly for second language learning and speech therapy purposes. In this situation, an instructor aids the subject to reach a targeted vocal tract configuration by showing them a live augmented visualization of the trajectories of (some of) the subject's articulators, alongside a targeted configuration.

Current research into how subjects respond to this training uses a variety of different visualizations: Katz et al. (2010) and Levitt et al. (2010) used a 'mouse-controlled drawing tool' to indicate target areas as circles on the screen, with the former displaying an 'image of [the] current tongue position', the latter displaying a 'tongue trace'. Suemitsu et al. (2013) displayed a mid-sagittal representation of the tongue surface as a spline between three sensors along the tongue, as well as showing a palate trace and lip coil positions and targets as circles. Katz and Mehta (2015) used a 3D avatar with a transparent face mesh, pink tongue rig, including colored shapes that lit when touched as targets.

For audiovisual feedback scenarios the optimal manner of presenting the stimuli has not yet been explicitly studied, but rather the experiments have reflected recent software developments. Meanwhile, different tools (Tiede, 2010; Ouni et al., 2012) have emerged as state of the art software for offline processing and visualization. The claim that subjects make gains in tongue gesture awareness only after a practice period with the visualization (Ouni, 2011) underlies the need for research into how EMA visualizations can best be presented to subjects in speech therapy or L2-learning settings.

The main inspiration for this work is the finding of Badin et al. (2010) that showing normally-obscured articulators (as opposed to a full face, with and without the tongue) has a positive effect on the identification of VCV stimuli. An established body of research already focuses on quantifying the intelligibility-benefit or realism of animated talking heads, ideally as compared to a video-realistic standard (Ouni et al., 2007; Cosker et al., 2005). However, as the articulators that researchers/teachers wish to present to their subjects in the aforementioned scenario are generally outside the line of sight, these evaluation methods cannot be directly applied to intra-oral visualizations. We aim to fill this gap by comparing commonly-used EMA visualizations to determine which is most intelligible,[1] hoping this may guide future research into the presentation of EMA data in a visual feedback setting.

[1] This word-identification task differs from the most common speech-training usage whereby a learner's attention is drawn to the difference between a live animation of their movements and some reference placement or movement.

Proceedings of the 14th Annual SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology, pages 87–92,
Berlin, Germany, August 11, 2016. ©2016 Association for Computational Linguistics

2 Method

In this experiment, animations of eighteen CVC English words were presented in silent conditions to participants of differing familiarity levels with vocal tract animations in an online survey; subjects were asked to identify the word in a forced-choice paradigm (a minimal pair of the prompt could also be chosen) and later give qualitative feedback about their experience speech-reading from the different systems.[2]

2.1 Participants

Participants were recruited through promotion on social media, university mailing lists, on the internet forum *Reddit* and on *Language Log*. In sum, 136 complete responses were collected, with three of these excluded for breaking the experiment over several days. We analyze the results of all 84 native English speakers. Participants had varying levels of previous exposure to vocal tract animations: of those analysed 43% had seen such animations before, 25% had no exposure, 25% had studied some linguistics but not seen such animations, and 6% considered themselves experts in the topic.

2.2 Stimuli

The prompts presented were nine minimal pairs of mono-syllabic CVC words spoken by a single British female speaker recorded for the study of Wieling et al. (2015).

Three of the pairs differed in the onset consonant, three in the vowel, and three in the coda consonant. Care was taken that the pairs had a significant difference in place or manner that would be visible in the EMA visualization.

In order to compare the animations, they were standardized as follows: a frontal view was presented on the left half of the screen, a mid-sagittal view with the lips to the left on the right half. No waveform or labeling information was displayed. Lip coils were green, tongue coils red and chin/incisor coils blue. Where surfaces were shown, lips were pink, and tongues were red. A palate trace, made using each tool's internal construction method, was displayed in black. A white or light grey background was used.

Onset	Nucleus	Coda
sad/bad	bet/bit	time/ties
mess/yes	mat/mitt	sum/sun
bale/tale	whale/wheel	maid/male

Table 1: Prompt minimal pairs, by location of difference.

The animations were produced as follows: **Dots with tails** were produced using functions from Mark Tiede's MVIEW package (Tiede, 2010), with an adapted video-production script for the standardizations mentioned above. **2D animations** were produced from VisArtico (Ouni et al., 2012), using the internal video-production processes. **3D animations** were produced using a simulated real-time animation of the data in Ematoblender (James, 2016), which manipulates an adapted facial rig from MakeHuman in the Blender Game Engine. See Figure 1 for examples of the three types of visualizations.

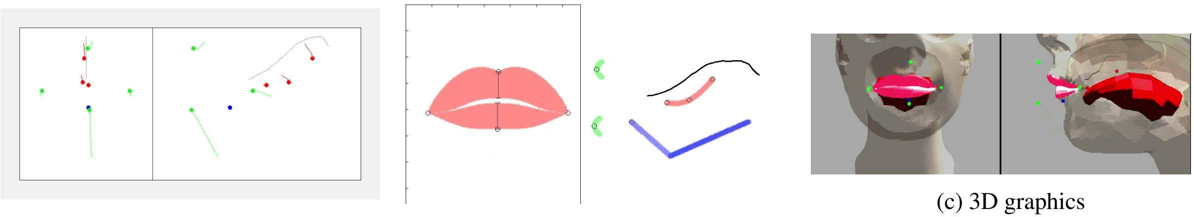
2.3 Procedure

This experiment was hosted on the platform SurveyGizmo. Firstly the EMA data was explained and participant background information was collected. This included information about previous exposure to linguistics studies and vocal tract visualizations. A brief training session followed, in which participants saw four prompts covering a wide range of onset and coda consonants in all three animation systems. They were free to play these animations as many times as they wished.

Subsequently, subjects were presented with two silent animations. The animations were either matching or non-matching (minimal pair) stimuli, which were displayed as HTML5 videos in web-friendly formats. They were controlled only by separate 'Play' buttons below each video. For each of these animations the subject was presented with four multiple choice options (one correct, one minimal pair, one randomly chosen pair, with the items and order retained across both questions). They were also asked to rate whether they believed the two stimuli to be the same word or not.

Upon submitting their answers, the subject was asked to view the videos again (as often as they liked) with sound, allowing them to check their answers and learn the mapping between animation and sound. The time that they spent viewing each prompt (for identification and after the an-

[2] The experimental design also collected data about whether subjects could perceive differences between the competing animation paradigms, for a separate research question.



(a) Dots with tails

(b) 2D graphics

(c) 3D graphics

Figure 1: Different animation paradigms tested.

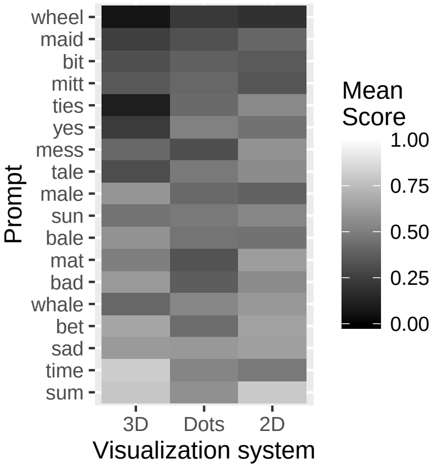


Figure 2: Prompt variability by animation type. Lighter colors indicate a better response.

	ID model
(Intercept)	$-0.26(0.14)$
SYSM	$-0.08(0.11)$
SYSV	$0.27(0.11)^{*}$
AIC	3261.90
Num. obs.	2486
Num. groups: RESPID	83
Num. groups: PROMPT	18

$^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$

Table 2: Coefficient, standard error and significance and of fixed effects for the mixed model of the identification dataset. 2D animations (SYSV) improve identification significantly over the baseline (3D animations). Table created with texreg (Leifeld, 2013).

swer was revealed) was also measured. After each three questions they were asked to rate their confidence at guessing the prompts' identities. Then after twelve questions they were asked to comment about their strategies. Finally, they could complete another six questions, or skip to the concluding qualitative questions.

3 Data Analysis

The prompt identification task yielded a binomial dataset based on the correctness of the identification. The random assignment of prompt pairs to system combinations led to an unbalanced dataset, which motivated the use of generalized linear mixed-effects regression models (GLMMs) for analysis (Bates et al., 2015). Random intercepts and slopes were included if they improved the model in a model comparison procedure.

In order to take into account the variability in subject responses, random intercepts for subject were included. Similarly, random intercepts were included for each prompt. The prompt variability was quite extensive and is visualized in Figure 2.

4 Results

The resulting model for the identification data included random intercepts for the subject, random intercepts for the prompt (with a random slope for the match-mismatched condition), and a fixed effect for the system, shown in Table 2. The 2D animation was significantly better-identified than the 3D animation. The Dots animation was slightly (but not significantly) less well-performing than the 3D animation.

Even within the most intelligible system (2D graphics), it is evident that there is much variability in how well participants are able to identify the various prompts (see Figure 2). A generalized logistic mixed-effects regression model was fitted to analyze the effects of onset and coda consonants and the nuclear vowel in the prompts.

When assessing the effect of either onset, coda or nucleus on how well people were able to detect the correct utterance, we found that the type of nucleus (i.e. the vowel) was most important. For example, whenever a stimulus contained the vowel /a/ its recognition was better than with a different vowel. In contrast, a stimulus with the vowel /i/ was much less well recognized. As the vowel necessitates greater movements of especially the lips than consonants, it makes sense that the type of vowel is an important predictor. Given that we only had a limited number of stimuli, including the onset or coda together with the nucleus did not help predict the recognition of the stimulus.

The hypothesized effect on the identification score of question number and time spent watching the videos (a learning effect was expected) was not borne out in the results. Though many subjects improved over time, others worsened, which could be attributed to fatigue or boredom during the long experiment. Similarly, including the subjects' previous experience with linguistics and vocal tract visualizations did not significantly improve the model.

5 Discussion

5.1 Identification strategies

The model's identification of the ease of interpreting 2D animations was reflected in participants' comments about the strategies they used for speech-reading. The frequency with which these strategies were mentioned is shown in Table 3.

Strategy	Frequency
Lip aperture/shape	71
Mimic the animation	56
Tongue placement/movement	48
Tongue-palate contact/distance	25
Knowledge of phonetics	21
Deduce using answer options	15
Tongue timing	7
Start/end position	5
Counting syllables/gestures	3
Vowel length	2
Visualize someone else	1

Table 3: Identification strategy frequency by number of mentions over all participants.

One participant (ID 1233) summed up the particular difficulty of the 'dots with tails' system succinctly: "In the ones with lips and tongue, I spoke each of the possible answers myself and tried to envision how closely my own lips and tongue resembled the videos. In the one with just dots, I was purely guessing."

5.2 Pitfalls of the 3D animation

Whereas it might seem somewhat surprising that the 3D animation did not result in (significantly) better recognition over the simplest representation (dots with tails), participants' comments highlight some possible causes.

Firstly, the colors of the lips and tongue were similar, which was especially problematic in the front view of this experiment. Though the color choices were made based on VisArtico's color scheme, the 2D animation avoids this problem by excluding the tongue from the frontal view.

Secondly, participants expressed that they would have liked to see teeth and a facial expression in the 3D animation. They also commented that they expected more lip-rolling movement. Indeed, seeing a more realistic avatar with these crucial elements missing may have been somewhat unnatural-looking.

Some linguistically-experienced participants also indicated that they expected a detailed 3D avatar to also indicate nasality, the place where the soft and hard palates meet, or 'what the throat is doing'. Unfortunately, this information is not available using EMA data.

Finally, many subjects commented that they found the 3D animation 'too-noisy' and preferred the 'clean' and 'clearer' 2D option.[3] Subjects' descriptions of their personal identification strategies indicates that they often used lip-reading strategies, and that this was easier in 2D where the lip shape was clear, and there was no difficulty with any color contrasts from the tongue. While the graphics quality of the 3D system was not as clear as for the other systems, the setup is similar to the 3D state of the art such as reported in Katz et al. (2014).[4]

5.3 Additional observations

Though the speaker and analyzed participants all identified themselves as English native speakers, two American participants noted that they

[3] Due to a combination of video capture technique and data streaming rate (the 3D system was recorded with real-time processing) the frame rate of the 3D system was lower than the other systems. Consequently, some participants also commented they wished for smoother 3D animations.

[4] The shapes of the tongue and lips in the 3D animation are controlled by internal constraints within the Blender Game Engine, and are dependent on the mesh shape. The performance of the 3D graphics could be improved by using a more-detailed facial rig and mesh and allowing a slower rendering (or using a faster game engine).

perceived the British speaker as having a foreign/German accent. Several participants mentioned that their main tactic was mimicking the speaker saying the answer options (and in doing so mimicking their interpretation of the speaker's accent), which they on occasion found difficult. This underlines the usefulness of using dialect-appropriate trajectories for the speech-reader.

In this experiment, all animations were based on EMA recordings from a single speaker in one recording session. In general usage however, the differing coil placement for each subject and recording session may also affect the identification ability. Other visualization methods (e.g., cine-radiography or MRI) give a high-dimensional picture of the vocal tract and avoid these problems. However, these technologies are not practical for real-time speech training due to their health-risk and cost, respectively. One strategy to compensate for this problem when creating the animations is to use photos of the coil placement during recording to manually specify the offset from the intended placement on the articulator. For example, VisArtico allows the user to specify whether the lip coils were placed close to or above/below the lip opening.

6 Conclusion

In sum, the simplicity and clarity of 2D graphical animations is preferable for subjects to identify silent animations of EMA data. The features of the most successful animation paradigm suggest that future EMA-animations should include both indications of lip and tongue surface shape. If used, 3D models should ensure that they provide clear and clean demonstrations, in which the edges of the articulators (particularly in the frontal view) can easily be distinguished.

References

Pierre Badin, Yuliya Tarabalka, Frédéric Elisei, and Gérard Bailly. 2010. Can you 'read' tongue movements? Evaluation of the contribution of tongue display to speech understanding. *Speech Communication*, 52(6):493–503.

Douglas Bates, Martin Mächler, Ben Bolker, and Steve Walker. 2015. Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, 67(1):1–48.

Blender. `https://www.blender.org/`. Accessed: 2016-04-14.

Darren Cosker, Susan Paddock, David Marshall, Paul L Rosin, and Simon Rushton. 2005. Toward perceptually realistic talking heads: Models, methods, and mcgurk. *ACM Transactions on Applied Perception (TAP)*, 2(3):270–285.

Kristy James. 2016. Watch your tongue and read my lips: a real-time, multi-modal visualisation of articulatory data. Master's thesis, Saarland University/University of Groningen.

William F Katz and Sonya Mehta. 2015. Visual feedback of tongue movement for novel speech sound learning. *Frontiers in human neuroscience*, 9.

William F Katz, Malcolm R McNeil, and Diane M Garst. 2010. Treating apraxia of speech (AOS) with EMA-supplied visual augmented feedback. *Aphasiology*, 24(6-8):826–837.

WF Katz, Thomas F Campbell, Jun Wang, Eric Farrar, J Coleman Eubanks, Arvind Balasubramanian, Balakrishnan Prabhakaran, and Rob Rennaker. 2014. Opti-speech: A real-time, 3D visual feedback system for speech training. In *Proc. Interspeech*.

r/languagelearning. `https://www.reddit.com/r/languagelearning/`. Accessed: 2016-04-14.

Philip Leifeld. 2013. texreg: Conversion of statistical model output in R to LaTeX and HTML tables. *Journal of Statistical Software*, 55(8):1–24.

June S Levitt and William F Katz. 2010. The effects of EMA-based augmented visual feedback on the English speakersácquisition of the Japanese flap: a perceptual study. *stroke*, 4:5.

Mark Liberman. 2016. Language Log. `http://languagelog.ldc.upenn.edu/nll/?p=24223`. Accessed: 2016-04-14.

MakeHuman Open Source tool for making 3D characters. `http://www.makehuman.org/download.php`. Accessed: 2016-02-09.

Slim Ouni, Michael M Cohen, Hope Ishak, and Dominic W Massaro. 2007. Visual contribution to speech perception: measuring the intelligibility of animated talking heads. *EURASIP Journal on Audio, Speech, and Music Processing*, 2007(1):3–3.

Slim Ouni, Loïc Mangeonjean, and Ingmar Steiner. 2012. VisArtico: a visualization tool for articulatory data. In *13th Annual Conference of the International Speech Communication Association-InterSpeech 2012*.

Slim Ouni. 2011. Tongue Gestures Awareness and Pronunciation Training. In ISCA, editor, *12th Annual Conference of the International Speech Communication Association - Interspeech 2011*, Florence, Italy, August. (accepted).

r/samplesize. `https://www.reddit.com/r/SampleSize`. Accessed: 2016-04-14.

SurveyGizmo. `http://www.surveygizmo.com/`. Accessed: 2015-11-13.

Atsuo Suemitsu, Takayuki Ito, and Mark Tiede. 2013. An electromagnetic articulography-based articulatory feedback approach to facilitate second language speech production learning. In *Proceedings of Meetings on Acoustics*, volume 19, page 060063. Acoustical Society of America.

Mark Tiede. 2010. MVIEW: Multi-channel visualization application for displaying dynamic sensor movements. *unpublished*.

Martijn Wieling, Pauline Veenstra, Patti Adank, Andrea Weber, and Mark Tiede. 2015. Comparing L1 and L2 speakers using articulography. In *Proceedings of the 18th International Congress of Phonetic Sciences*. University of Glasgow, August.