Visual Learning 2: Pronunciation app using ultrasound, video, and MRI

Kyori Suzuki, Ian Wilson, Hayato Watanabe

CLR Phonetics Lab, University of Aizu, Japan

m5211115@u-aizu.ac.jp, wilson@u-aizu.ac.jp, s1230063@u-aizu.ac.jp

Abstract

We demonstrate *Visual Learning 2*, an English pronunciation app for second-language (L2) learners and phonetics students. This iOS app links together audio, front and side video, MRI and ultrasound movies of a native speaker reading a phonetically balanced text. Users can watch and shadow front and side video overlaid with an ultrasound tongue movie. They are able to play the video at three speeds and start the video from any word by tapping on it, with a choice of display in either English or IPA. Users can record their own audio/video and play it back in sync with the model for comparison.

Index Terms: pronunciation, mobile application, L2 learning, ultrasound, video, MRI

1. Introduction

Although many language learners wish to improve their foreign language pronunciation, they lack effective apps to help do so. Most commercial apps for pronunciation evaluation and training focus on only the acoustic signal. Articulatory feedback has been shown to be helpful to pronunciation learners [1], but systems that include it tend to require complex equipment [2]. However, few simple apps model visual movements of the lips, tongue, and jaw [3].

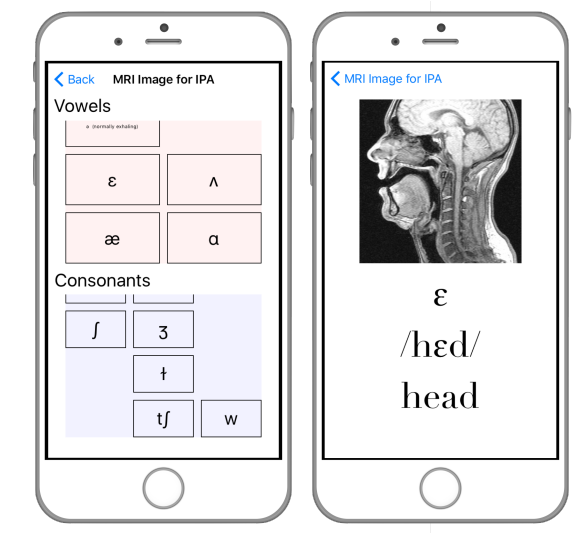
In 2016, we presented and got feedback on *Visual Learning*, a new pronunciation app for the iOS platform [4]. The app's potential users are L2 pronunciation learners, phonetics students and teachers, and speech language pathologists. In version 1, we had built an original AVPlayer enabling users to see an English native speaker's real (non-animated) tongue movements with ultrasound, real lip and jaw movements with video, and clear tongue postures with MRI. The training text was the English paragraph from the Speech Accent Archive [5], which has audio samples available online by over 2,350 speakers native in a range of more than 300 languages and dialects.

Based on feedback received, we have improved the app, and now demonstrate *Visual Learning 2*. Added features include the ability to record user video (focusing on the mouth), play back in sync with the model video, improved sound quality, and use of a new phonetically balanced passage (the 'Wolf' passage) that has a greater range of phonetic environments and naturally occurring minimal pairs [6]. We are now testing how much this system helps people learning L2 pronunciation, as they shadow the audiovisual speech of a native speaker [7].

2. Features of the app

One feature of *Visual Learning 2* is the ability to display a table of all International Phonetic Alphabet (IPA) phonemes in the Wolf passage (Figure 1(a)), and corresponding midsagittal MRI images of the same speaker as the video images (Figure 1(b)).

Figure 2 shows four screenshots of this app. The Sentence screen in (a) allows users to focus on a single phrase or the whole 'Wolf' passage. Phrases can be displayed in English or in IPA. When the user taps a phrase, the screen moves to (b).



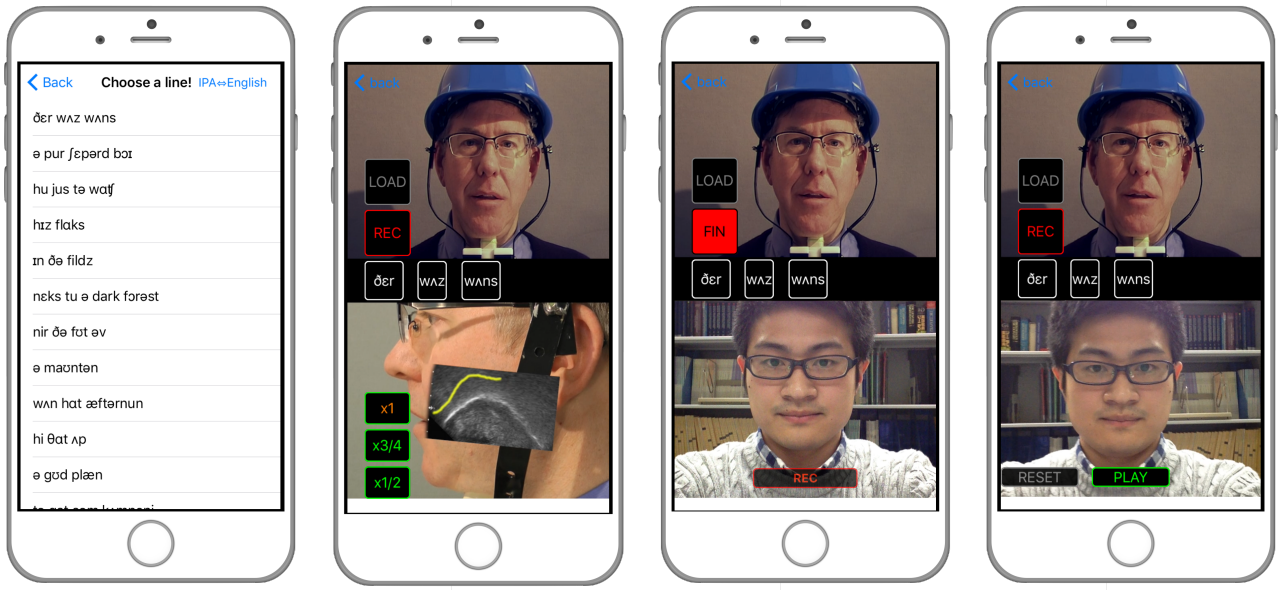
(a) IPA phonemes (b) Midsagittal MRI

Figure 1: *Screenshots of the MRI feature*

Figure 2(b) shows front and side videos of a native English speaker. The side video is overlaid with an ultrasound movie of the tongue's surface (white line) moving in the mouth and shows the palate (yellow line determined via MRI overlay). Buttons for each word of the user-chosen phrase appear horizontally in the middle and can be tapped on to start the video from there. Two slow-motion speeds are available. If the user taps the record button, the screen moves to Figure 2(c).

In Figure 2(c), the user can record his/her face and voice, while simultaneously playing the model video and shadowing the top half of the display. Wearing earphones ensures that only the user's voice is recorded. After this, the screen moves to (d), where the user can then load the recorded video from his/her video library. The user can then play both videos (top and bottom) simultaneously for comparison.

The native-speaker model data was recorded using a variety of equipment. MRI data was obtained at ATR-BAIC (Kyoto, Japan) using a Shimadzu-Marconi Magnex Eclipse 1.5T MRI. Front and side videos were recorded using a Victor GZ-HD7-S and a Panasonic HDC-TM750 videocamera, respectively, along with halogen lighting. Sound was recorded using a Korg MR-1000 mobile recorder with a DPA 4080 miniature cardioid mic. A Toshiba Famio 8 ultrasound with a SSA-530A probe was used to record tongue movies. A clapper was used to make subsequent signal alignment easier in Final Cut Pro 10.3.2.



(a) Sentence screen (b) Movie screen (c) Recording screen (d) Playback screen

Figure 2: *Screenshots of Visual Learning 2*

3. Coding issues in iOS

One coding challenge was changing movie speed. Initially, when a user changed playback speed, there was a finish-time lag. Now, when tapping a change-speed button, a timer is set and the time interval is divided by the rate, thus changing the duration correctly. Thus, a user can successfully start and finish playing the video at any word, at any speed.

Another challenge was making useful features that include some elements about words by using *enum* function. *Enum* can efficiently define an integer constant with names for that assigned variable. By doing this, this app allows changing and adding words easily.

When recording a movie with iOS, the default dimensions match the display, so on an iPhone, for example, the movie is longer in one dimension. However, by matching only the center y-coordinate of the actual recording display to the center y-coordinate of the desired recording area in a view, the user's face appears the same size as the model's face without compressing the face to fit the new half-screen view.

Finally, the reason the user can record and play back while playing the model movie is that recording view and playback view do not interface with the original player. Also, because we built these as modules, the app has good scalability.

4. Intended use and future work

The L2 user first chooses a training phrase or 'all sentences', watching and listening to how to pronounce the phrase from the movie while checking MRI images of individual phonemes or changing the video speed. Next, the user shadows the video, an effective learning technique where you practice imitating the speaker while she or he is talking [7], and finally plays back and compares the recorded video in sync with the model.

After making this app freely available on the App Store, we will get usability feedback and research the app's actual effect on learners' pronunciation. Future updates will include features such as automatic image comparison of user's versus model's lip and mouth shape, automated feedback to user regarding possible tongue position errors based on audio, and optional uploading of audiovisual data to a university server for further analysis.

5. Acknowledgements

We thank N. Horiguchi, Y. Iguro, and R. Omori for motivation & assistance, T. Hozumi (U of Aizu Geek Dojo) for designing and 3D printing the ultrasound probe holder, and attendees for feedback at the 5th Joint Meeting of the ASA/ASJ. Parts of this work were supported by Japan Society for the Promotion of Science kakenhi grants #19520355, #23520467, and #25370444.

6. References

[1] A. Suemitsu, J. Dang, T. Ito, and M. Tiede, "A real-time articulatory visual feedback approach with target presentation for second language pronunciation learning," *J. Acoust. Soc. Am.*, vol. 138, pp. EL382–EL387, 2015.

[2] W. Katz, T. 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," in *Proc. of Interspeech*, 2014, pp. 1174–1178.

[3] N. Horiguchi and I. Wilson, "Design of an interactive GUI for pronunciation evaluation and training," *Proc. of the 12th International Conference on Humans and Computers*, pp. 225–229, 2009.

[4] K. Suzuki and I. Wilson, "Development of a visual app for improving learner's pronunciation with ultrasound and the Speech Accent Archive," *J. Acoust. Soc. Am.*, vol. 140, p. 3343, 2016.

[5] Speech Accent Archive. [Online]. Available: http://accent.gmu.edu/

[6] D. Deterding, "The North Wind versus a Wolf: short texts for the description and measurement of English pronunciation," *J. Int. Phon. Assoc.*, vol. 36, pp. 187–196, 2006.

[7] A. Rojczyk, "Phonetic imitation of L2 vowels in a rapid shadowing task," in *Proc. of the 4th PSLLT Conference*, 2013, pp. 66–76.