COMPUTER-VISION-AIDED LIP MOVEMENT CORRECTION TO IMPROVE ENGLISH PRONUNCIATION

Ms. Shuang Wei, Purdue University, West Lafayette

Shuang Wei is a Ph.D. student in the department of Computer Graphics Technology, Purdue University. She received her Master of Science degree from the same major and a Bachelor degree in digital media from HIT University (China). Her research focuses on multimedia education, information visualization, and human computer interaction.

Dr. Yingjie Chen, Purdue University, West Lafayette

Dr. Yingjie Chen is an assistant professor in the Department of Computer Graphics Technology of Purdue University. He received his Ph.D. degree in the areas of human-computer interaction, information visualization, and visual analytics from the School of Interaction Arts and Technology at Simon Fraser University (SFU) in Canada. He earned the Bachelor degree of Engineering from the Tsinghua University in China, and a Master of Science degree in Information Technology from SFU. His research covers interdisciplinary domains of information visualization, visual analytics, digital media, and human computer interaction. He seeks to design, model, and construct new forms of interaction in visualization and system design, by which the system can minimize its influence on design and analysis, and become a true free extension of human's brain and hand.

Dr. Tim McGraw, Purdue University
April Ginther, Purdue University

Associate Professor, Second Language Studies, Associate Professor Linguistics, Director Oral English Proficiency Program, Co-Editor Language Testing

COMPUTER-VISION-AIDED LIP MOVEMENT CORRECTION TO IMPROVE ENGLISH PRONUNCIATION

Introduction

Mastering English is an essential part of the study program for international students in the United States. Although self-repetition[1] and self-evaluation [2] are important methods to improve English pronunciation, self-evaluation is very hard because people often can't realize the mistakes they make. Computer-Assisted Pronunciation Training (CAPT) software provides opportunities for student self-repetitions and self-evaluations. To help users evaluate their pronunciation, a variety of feedback is provided to users, such as record comparison, spectrum, and shape-movement comparisons of the mouth.

The lips are one of the pronunciation mechanisms. Lip movement is an important factor that influences the pronunciation of learners.[3] It is helpful for learners to get visual feedback about their mouth-shape movements. Through feedback, learners can directly find where the mouth movement is wrong and learn how to correct it. Arai & Oda[4] used 3D models to help users understand the correct mouth-lip shapes in their research. The evaluations showed that visual lip-movement feedback helped to improve the pronunciation of deaf people for single sounds.

With computer-vision technology, an English learner can compare his own mouth-lip movements with standard mouth movements while speaking. Our research was designed to answer the following research questions:

Can lip movement feedback provided by computer-vision-aided English language training improve the pronunciation of English as a Second Language (ESL) learners?

To what extent can the method help ESL learners improve their pronunciation of English?

In this paper, we demonstrate our solution of using computer-vision-aided lip movement correction to help students learn English pronunciation. We also evaluate the effectiveness and efficiency of the method in helping ESL learners improve their pronunciation.

Literature Review

English pronunciation is crucial to English-language learners because "pronunciation is the language feature that most readily identifies speakers as nonnative."[5] ESL learners who have pronunciation weaknesses can be hard to understand, and they can feel embarrassment in conversation. Pronunciation is so important that "in the International Phonetic Association's declaration of principles of second-language teaching, the spoken language is held to be primary, and training in phonetics is important for both teachers and learners."[6]

English Phonetics

All languages have two categories of sounds: vowels and consonants. According to Delahunty and Garvey, the qualities of vowel sounds are determined by the position of the tongue and "the tension of the muscles and the configuration of the lips." [7] So the pronunciation of vowel sounds is influenced by the tongue position and the lips.

The way to pronounce consonants relates to vocal cords, teeth, lips, and other factors. The state of the vocal cords influences the articulation of consonants. When vocal folds are relaxed, the flow of air passes freely through the glottis, so the sound is "voiceless." When vocal folds vibrate, the sound is "voiced." [8]

The point of articulation also makes a difference. Interdentals (θ, ð) are made by putting the tongue between the front teeth; bilabial sounds (p, b, and m) are made by bringing both lips closer together; and labiodental consonants (f, v) are made with the lower lip against the upper front teeth. [8] To improve the quality of pronunciation, learners should notice the position changes of jaw, lip, teeth, and part of the tongue.

Computer-Assisted Pronunciation Training

Fraser stated that a good way for educators to teach English pronunciation is by "having a suitable curriculum, being student-centered, helping learners become self-reliant, giving opportunities to practice, and knowing what's best." [9] Of these five principles, being student-centered and giving opportunities to practice are very hard to complete during courses because instructional time is limited, and pronunciation correction requires a large amount of time. [2] Instructors must leave some of these tasks to students and let them practice by themselves. (Computer-Assisted Pronunciation Training) CAPT applications can help students' self-learning by providing repeated lessons as often as needed. [10] Moreover, textual, visual, and audio feedback provided by CAPT applications may help users judge whether a task is completed correctly.[11] Researchers have found positive effects in the pronunciation achievement of students who use CAPT software along with traditional methods.[12]

In 1994, Inouye et al. published the patent "Method for teaching spoken English using mouth position characters."[13] In the patent, they described a speech training system that allowed learners to listen to a phrase, reference with the mouth-position characters, and speak along with the video. By using a mirror, students could compare their mouth-shape movements with the standard pronunciation movement and make needed improvements.

Today, Inouye's method is still used, but more instructional elements have been added. An example is SAUNDZ, which is available in the Apple iTunes Store. Using the application, learners can look at a text description, listen to the standard pronunciation, and imitate the visual animation. Furthermore, the application can record the pronunciation of users and "match" it with standard pronunciation. Learners can listen to the audio and find their mistakes. The application provides rich instruction and effective audio feedback. As stated in the slogan of the application, users can "hear it, see it, record it, and compare it!"

Automatic Speech Recognition (ASR) technology can analyze a user's speech in real time and transcribe the spoken language into text or spectrograms.[14] Developers put a spectrogram in an English pronunciation training application to give users visual feedback; thus learners can know which part of the word is mispronounced. "Tell Me More English" is a typical example. In the application, the pronunciation of the user is recorded and converted into a spectrogram. Users can visually compare their pronunciation with the standard pronunciation. This kind of application is helpful for tone and prosodic training, but "segmental errors cannot be shown clearly in this kind of application."[15]

Monitoring mouth-shape movement is another way to provide visual feedback to learners. Oda and Ichinose developed an application called "Lip Reading AI" in 2007.[16] The system allows users to look at their mouth-shape movements and compare them with standard movements. To make the application more efficient, in 2012 Arai and Oda integrated computer graphic (CG) animation into the application.[4] They created the user's 3D face model in advance. The model was then used to show the standard pronunciation mouth-shape movements. In this way, it was easier for users to compare because individual differences were eliminated. The application was evaluated in a study in which eight kindergarten boys and girls participated. "After pronunciation practice, children's pronunciation is improved by 3% to 9% from before pronunciation practice. This result implies that their pronunciation is certainly improved." [4] The approach is helpful to improve single sounds. However, for a word or a sentence, the mouth shape will change too quickly, and it will be difficult for the user to compare two videos side by side. Besides that, a 3D face model needs to be prepared for each user, which limits the generalization of the application. More work needs to be done to provide valid mouth-shape-movement comparison and real-time visual feedback.

Computer-Vision-Aided Lip-Movement Correction System

We hypothesized that by using computer-vision technology to show the difference of the mouth-shape movements of the user and the standard mouth-shape movements, learners could discern the differences, and work to improve their pronunciation. To approve the hypothesis, a prototype system was developed to test different ways of training and evaluate the effectiveness of the method.

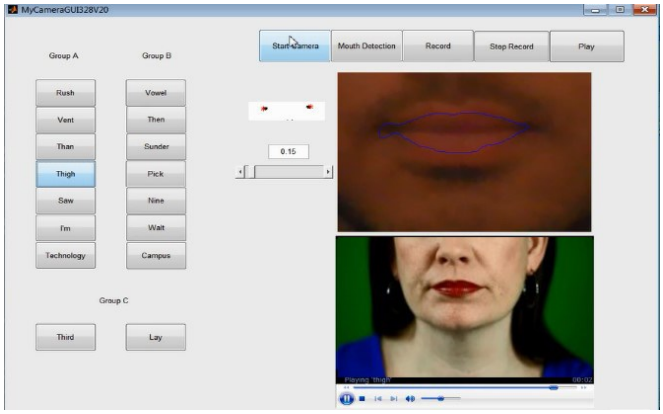

Figure 1: The main interface of the system

The system (Fig. 1) keeps a repository of practice words and their pronunciation videos. Since this system is a prototype to evaluate the proposed pronunciation training method, we collected only 16 words in the repository. The principles were used in choosing these words: they include obvious mouth-shape-movement changes, and they have some challenges for the participants.

To use this system, the user chooses the word that he/she wants to practice. The player window in the lower right-hand corner shows the corresponding standard pronunciation video to the user (Fig. 1).The user can then compare his/her own mouth-shape movement with the standard one by interacting with the system following these steps: start the camera, detect self-mouth corners, record self-pronunciation, and play the mouth-shape movement comparing the video.

To compare the user's mouth shape with the standard one, the mouth-shape movement video is created by overlaying the standard mouth-shape movement contour (blue lines) onto the recorded pronunciation video. The user may easily compare and find his own mistakes through the help of blue lines (Fig. 2).

Figure 2: Standard mouth-shape-movement contour and comparing video

There were some technical challenges in the development of the prototype, for example, extracting standard-pronunciation video information effectively and detecting mouth corners accurately.

The original idea was to directly overlay the standard pronunciation video onto the record video to show more information, such as the position of lips, teeth, and tongue (Fig. 3).

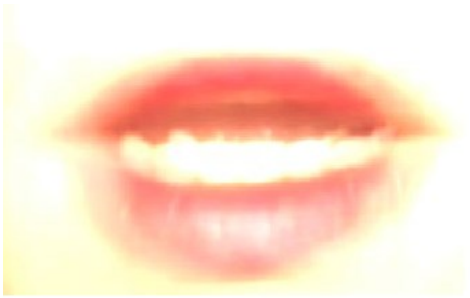

Figure 3: Overlay of the pronunciation video and recorded video

However, too much information and similar colors made it hard to distinguish between the user's mouth and the standard mouth. Directly overlaying two videos is not a good way to visualize the mouth-shape-movement information. To make the comparison video more efficient, standard lip movement contour was extracted and overlaid with the recorded video by using computer vision technology.

"For computer vision, the most significant and fundamental technical limitation is its robustness in the face of changing environmental conditions." [17] Different lighting, different users, and even different skin tones can lead to different results. To get an accurate and smooth contour, we chose a quiet studio with good lighting conditions and let an English-pronunciation trainer wear lipstick to complete the recording of standard pronunciation. Lighting is very important to the extraction of contour, and lipstick can make the lip contour more easily distinguished. After recording, the frames of the standard pronunciation were converted into binary images to find the mouth region, then edges of the mouth were detected, and finally any holes left over in the center of the mouth contour were filled (Fig. 4).

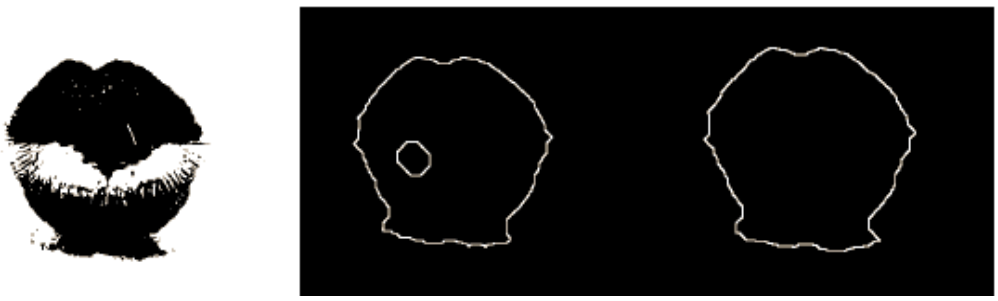

Figure 4: The extraction process of standard mouth-shaped-movement video

To match the standard pronunciation video and the recorded video, it is necessary to know the accurate position and scale of both the professional pronunciation trainer's mouth and the user's mouth. Because mouth corners are easily detectable feature points, they were selected as reference points to determine the mouth's position. The distance between the two mouth corners determines the scale of a mouth.

Since the standard mouth-shape movement contour was extracted, it's easy to extract mouth corners of the English trainer (left and right end points). The problem was to detect the user's mouth corners in real time. To do this, the system first finds the user's face. Then the lower third is recognized as the mouth region, which is also the detection region. Because the detection area is small, it is easier to obtain accurate mouth corners. For greater accuracy, adjusting the value of the threshold parameter is necessary to make the system effective for different people in different lighting conditions. The threshold in the system controls how the gray-scale image is converted into a binary image. The value of every pixel in a gray-scale image is between 0 and 255. When the image is converted into a binary image, the value of every pixel changes to 0 or 1. If the value of a pixel is higher than the threshold, it is 1 (white). If it is lower, it is 0 (black). Normally, the user's mouth corners are the darkest points. When the threshold rises above 0, the corners change to black before any other points. By adjusting the threshold, corners can be more easily detected. The standard pronunciation video is overlaid onto the user's pronunciation video after the corners are detected.

Although there were some challenges in developing the prototype, it basically realized the initial design and implemented necessary functions that could be used to evaluate the proposed pronunciation training method.

Evaluation

An evaluation was conducted based on the developed prototype system to evaluate the effect of the mouth-shape-matching approach to improve ESL learners' pronunciation.

To eliminate the influence of English proficiency differences between different participants, a pretest-posttest control group design was used to evaluate the effect of the approach by comparing it with the standard mouth-shape movement video. The difference between posttest and pretest is called the "gain score". In the experimental group the gain score is the result of using the system to practice word pronunciations. In the control group, the gain score is the result of using the standard video to practice word pronunciation. By analyzing gain scores, we can find out if the pronunciation for each student has improved.

To increase statistical power, we conducted a within-subjects design, which means that a participant took part in the control group and the experimental group at the same time. To accomplish this design, we provided two groups of words, A and B. Participants randomly used one treatment to practice one group of word pronunciation. However, a main weakness here is that the experience in one group may influence a participant's performance in the other. To decrease the influence of this weakness, the sequence of using treatment was different for different people; so there were four cases (Table 1).

Table 1: Different treatment combinations for evaluation

	First Round	Second Round
Case 1	Group A + System	Group B+ Video
Case 2	Group A + Video	Group B + System
Case 3	Group B + System	Group A + Video
Case 4	Group A + Video	Group B + System

By using within subjects design, both the control group and the experimental group could have many subjects, thus increasing the power of the experiment.

Twenty ESL learners from different majors, countries, and age groups were recruited. Each participant was randomly assigned to an experiment combination. As an example, for the first round: using video to practice Group A's words; for the second round: using the prototype to practice Group B's words. After the experiment, the subjects were required to take a survey. Their subjective opinions of the system, which is qualitative data, were collected.

To determine whether there was significant improvement in the pronunciation of participants, video data were evaluated and converted to quantitative data. To enforce investigator triangulation, two linguistics students who were native English speakers were recruited to help evaluate the videos. A five-point rating scale [18] was used in the evaluation.

1. Very strong foreign accent: definitely nonnative.
2. Strong foreign accent.
3. Noticeable foreign accent.
4. Slight foreign accent.
5. No foreign accent at all: definitely native.

Data Analysis

The research question is whether the method is able to improve the pronunciation of ESL learners; in other words, whether the mean of posttest pronunciation is higher than the mean of pretest pronunciation. The quantitative data from the evaluation show that the computer vision-

aided lip movement correction training can improve the pronunciation of participants, and for some specific words, the method can significantly improve the pronunciation. Moreover, the qualitative data show that participants agreed with the effectiveness of computer-vision-aided lip-movement-correction training.

Two evaluators graded the pronunciation thus every word had two scores. Each word's average score was calculated first. The average score of each word was then added to calculate the mean of the entire group of words by dividing it by the number of words in the group. The group mean is also the participant's pronunciation mean.

After the participant's pretest and posttest pronunciation means were calculated, a paired two-sample t-test for the mean was used to test the hypothesis. The result indicated there was improvement in the participants' pronunciation ($\mu_{post} = 3.686$; $\mu_{pre} = 3.589$). However, the improvement was not significant ($p = 0.767$).

But one thing attracted our attention, notably that the pronunciation scores of the words "vowel" and "I'm" were significantly improved after using the system ($p_{vowel} = 0.9689$; $p_{I'm} = 0.9876$). For the control group, however, using the standard video alone did not significantly improve the pronunciation of these two words ($p_{vowel} = 0.8688$, $p_{I'm} = N/A$).

Participant opinion data on the system were also analyzed. Figure 5 shows their opinions on the helpfulness of the system.

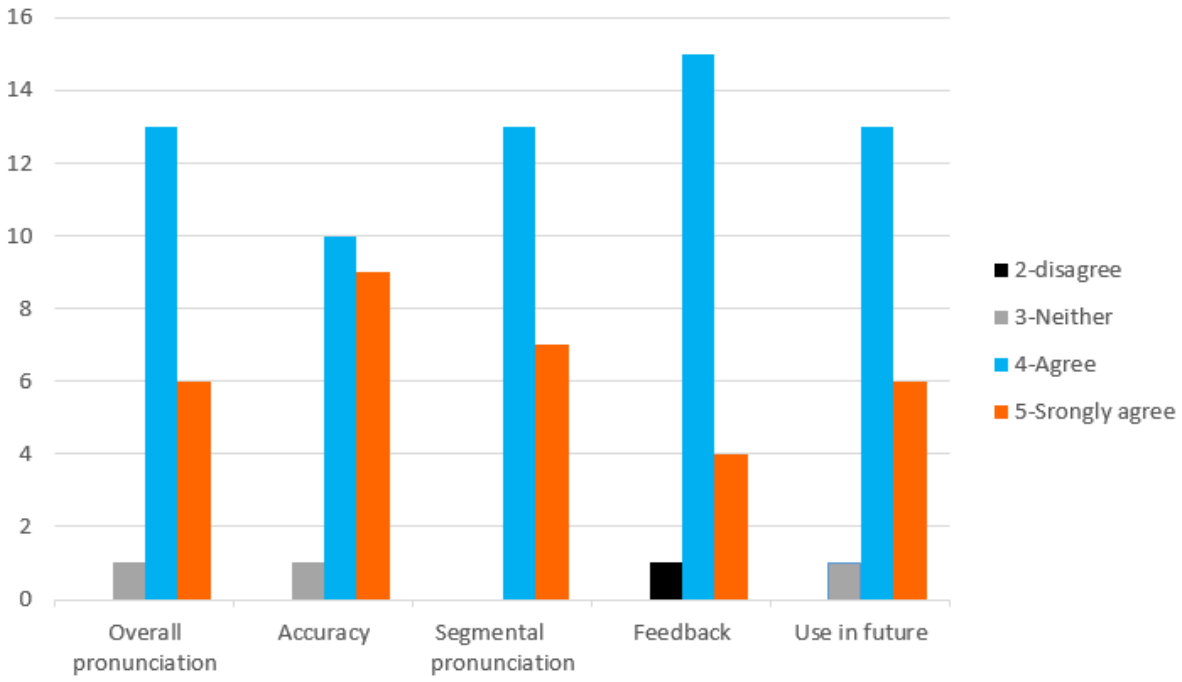

Figure 5: Opinions on the helpfulness of the system collected from evaluation

More than 95% of the participants agreed that the system improved their pronunciation accuracy and the segmental pronunciation features. They thought that the mouth-shape movement feedback was helpful, and they would like to use this system in the future.

At the same time, much advice about the system was given. Thirty percent of the participants pointed out that there were too many operation steps to use the system. This distracted the learners. Twenty percent of the participants thought the system could be improved if mouth detection was more accurate. More suggestions were made, such as adding sound analysis to the system, giving verbal suggestions to the user, and providing both male and female pronunciation videos.

The data analyses result partially support the hypothesis that computer-vision-aided lip-movement correction can improve the pronunciation of ESL learners. After using the system to practice English pronunciation, the pronunciation is significantly improved for some words (vowel, I'm). However, for the other words there is no significant improvement. Therefore this approach cannot be generalized to improve pronunciation of all English words in general.

Discussion

Why did only the pronunciation of "vowel" and "I'm" show significant improvement while the pronunciation improvement of other words was limited? Two main reasons were considered.

The first reason is the larger difference in the way vowel is produced. But in pronouncing the other words, "then" and "than," we find a subtle difference in the opening of the mouth, which is a phonological distinction between mid- and low-level vowels. Because most nonnative speakers will not have this distinction in their vowel inventories, they may not have the capacity to hear or see the difference in the opening of the mouth. Another discrepancy is that "then" and "than" were put into two separate groups in the test, thus making it even harder for participants to notice the difference. But the words "I'm" and "vowel" contain diphthongs, which require motion during the production of the sound. They require continuous closing of the mouth and are followed by a bilabial stop (/m/) and lateral liquid (/l/). The significant improvement in the pronunciation of these two words may be due to better production or to watching the slow and continuous rise of the oral tract.

Another reason is that for some words, the wrong pronunciation is not related to the mouth-shape movement, as in "rush". The evaluator noticed that some participant pronunciations sounded incorrect even when the mouth-shape movement was correct. So that may be the reason why the method could not improve every word's pronunciation.

Conclusion and Future Work

In this research, we tried to help English learners to correct their pronunciations by comparing their lip movement with standard-pronunciation mouth movements using computer vision technology. Evaluation shows this approach can improve the pronunciation of some words, but has limited effectiveness on other words. Research in English phonetics shows that English learning is a complex problem involving coordination of different parts, including vocal cords, teeth, tongue, and lips. Apparently our approach is not a one-stop solution, but it can serve as a valuable basis for future work.

Some similar phonetic sounds may also be improved by using this method, for example, [ɛ] and [æ]. These similar phonetic sounds with slight mouth shape differences could be tested together using our method to compare, practice, and improve pronunciation. This method could also be combined with audio wave analysis to provide efficient feedback to learners, and a full functional application for smartphones could be developed to help ESL learners. We would like to generalize the idea and believe that this method could help many ESL learners struggling to improve their English pronunciation.

Bibliography

1. Derwing, T. M., Rossiter, M. J. & Munro, M. J. Teaching Native Speakers to Listen to Foreign-accented Speech. *J. Multiling. Multicult. Dev.* **23,** 245–259 (2002).
2. Hismanoglu, M. An Investigation of Pronunciation Learning Strategies of Advanced EFL Learners. *Hacet. Univ. J. Educ.* **43,** 246–257 (2012).
3. Dobrovolsky, M. & Katamba, F. Phonology: the function and patterning of sounds. *Contemp. Linguist. Introd. Essex Addison Wesley Longman Ltd.* (1996).
4. Arai, K. & Oda, M. Effects of Pronunciation Practice System Based on Personalized CG Animations of Mouth Movement Model. *Int. J. Adv. Comput. Sci. Appl.* **3,** 125–130 (2012).
5. Goodwin, J. Teaching pronunciation. *Teach. Engl. Second Foreign Lang.* 117–137 (2001).
6. Seidlhofer, B. in *The Cambridge Guide to Teaching English to Speakers of Other Languages* (eds. Carter, R. & Nunan, D.) (Cambridge University Press, 2001). at <http://dx.doi.org/10.1017/CBO9780511667206.009>
7. Delahunty, G. P. & Garvey, J. J. *The English Language: From Sound to Sense*. (WAC Clearinghouse, 2010).
8. O'Grady, W. D., Dobrovolsky, M. & Katamba, F. *Contemporary linguistics: an introduction*. (Longman, 1996).
9. Fraser, H. Teaching pronunciation: A handbook for teachers and trainers. (2001). at <http://helenfraser.com.au/downloads/HF%20Handbook.pdf>
10. Dina, A.-T. & Ciornei, S.-I. The Advantages and Disadvantages of Computer Assisted Language Learning and Teaching for Foreign Languages. *Procedia-Soc. Behav. Sci.* **76,** 248–252 (2013).
11. Kim, I.-S. Automatic Speech Recognition: Reliability and Pedagogical Implications for Teaching Pronunciation. *J. Educ. Technol. Soc.* **9,** 322–334 (2006).
12. Talebi, F. & Teimoury, N. The Effect of Computer- assisted Language Learning on Improving EFL Learners' Pronunciation Ability. *World J. Engl. Lang.* **3,** (2013).
13. Inouye, K. K., Sheres, S. C. & Inouye, L. M. Method for teaching spoken English using mouth position characters. (1994).
14. Stuckless, R. Developments in real-time speech-to-text communication for people with impaired hearing. *Commun. Access People Hear. Loss* 197–226 (1994).
15. Hansen, T. K. Computer assisted pronunciation training: The four'K's of feedback. *Curr. Dev. Technol.-Assist. Educ.* 342–346 (2006).
16. Oda, M., Ichinose, S. & Oda, S. Development of a Pronunciation Practice CAI System Based on Lip Reading Techniques for Deaf Children. (2007).
17. Guan, A., Bayless, S. H. & Neelakantan, R. *Connected Vehicle Insights, Trends in Computer Vision: An Overview of vision-based data acquisition and processing technology and its potential for the transportation sector*. (The Intelligent Transportation Society of America. Technology Scan Series, 2011).
18. Bongaerts, T., Van Summeren, C., Planken, B. & Schils, E. Age and ultimate attainment in the pronunciation of a foreign language. *Stud. Second Lang. Acquis.* **19,** 447–465 (1997).