Visual Approach to Speech Sounds
Haruko Miyakoda
1 Department of English, Tsuda College, Tokyo, Japan
miyakoda@tsuda.ac.jp

Abstract
Many people often struggle to master the pronunciation of foreign languages without much success. One of the reasons why L2 learners are not successful is because teaching pronunciation in the classroom is usually marginalized. With the advent of computers, this problem may partially have been overcome, due to the fact that many different types of systems and software for autonomous learning have been developed, allowing learners to improve their pronunciation skills outside the classroom. However, there are few, if any, systems and software that can present a form of visual feedback that allows learners to actually understand what their problems are.

In this paper, we present the auditory-visual pronunciation system that we have developed. One of the key features of this system is that it employs easy-to-understand visuals of the speech organ that can be seen from different angles. Furthermore, movement of the speech organs can freely be adjusted by the instructors so that the learner’s movements (especially the deviant) can be highlighted by comparing them with those of the model samples.

Index Terms: pronunciation system, speech organs, visual feedback

1. Introduction
There is general agreement in the literature that teaching pronunciation is given the least attention in many English language classrooms [1]. One of the reasons why pronunciation is marginalized in this matter is because most instructors think that the goal of phonological instruction in the classrooms should be attaining reasonable intelligibility, rather than native-like pronunciation (e.g. [2], [3]). In other words, instructors would rather spend time dealing with other aspects of language learning such as grammar since advancement in pronunciation skills can just be a waste of time. This holds true especially for late L2 learners. For example, researchers such as Scovel have argued that the so-called critical period for the acquisition of L2 pronunciation exists, and that learners who start to learn a new language after a certain age will never be able to attain native-like pronunciation [4]. However, it is also true that pronunciation is by far the most perceptible aspect in distinguishing a non-native speaker from a native one, and major difficulties in pronunciation often result in the learners facing difficulty in finding employment [1] (or, to put it the other way around, better opportunities lie ahead for those who have acquired better pronunciation skills). This may be why pronunciation has recently regained importance with some university centers, and institutions such as the British Council are also actively promoting the teaching of pronunciation skills in their business courses [5, 6].

2. The role of visuals in language learning: focus on vocabulary learning
In developing a pronunciation system, one of the first decisions that need to be made is what kind of information to include in the system. Of course, no one can deny the use of auditory data for this type of system, but our main concern was whether in addition to auditory data, was there the need to take visual data into consideration? In order find out the role that visual data plays in language learning, we conducted an experiment that compared the effectiveness of different types of learning materials in vocabulary learning.

In the field of vocabulary learning, many different results have been obtained on which factor plays a significant role. For example, we can find many studies that support the effectiveness of visual factors, but when the learning effects of movies and still images were compared, movies were favored over the latter in some studies [7], while others claim the effectiveness of still images [8]. Since incompatible results based on different experiments have been reported in the literature, we conducted our own experiment by designing and developing four systems.

In order to find out what factor leads to effective learning, we conducted an experiment that compared vocabulary learning based on the following four methods: 1) learning the words with game-oriented activity (System 1); 2) learning the words within contexts (System 2); 3) learning the words with their pronunciations (System 3); and 4) learning the words with their image data (System 4) [9], [10]. We compared the effectiveness of the systems by conducting a vocabulary memorization experiment that consisted of three tests over 11 weeks. The first test was carried out just after the exercise, the second one after 2 to 3 weeks, and the third one after 10 to 11 weeks. 11 undergraduate students attending a university in Tokyo participated in the experiment.

![Figure 1. Comparison of test results of different methods of learning](image-url)
obtained for the traditional so-called “paper-and-pencil” learning method. The results demonstrate that learning by System 4 (i.e. the system using visual data) is the most effective in the long run.

We also conducted six patterns of the t-tests between the system scores of each pair in order to determine whether a significant difference can be observed. The result indicated that the learning effect is highest when both auditory and visual information are employed in the learning process. Based on this finding, we decided that employing these two types of information in our pronunciation system would also lead to better understanding of the pronunciation of foreign languages.

3. The visual element in pronunciation

The findings obtained from our experiment on vocabulary learning seem to support the viewpoint that visual data does indeed play a role in language learning. In this section, we briefly review some of the research findings concerning the use of visual element in pronunciation.

Research in the field of phonology has long been dominated by a focus on the auditory aspect. However, in actual face-to-face communication, a significant source of information about the sounds a speaker produces comes from visual cues such as lip movements [11, 12]. There are some studies that have reported that the information value of visual cues can be improved with training. In one study, hearing-impaired adults were trained in visual consonant recognition. After a total of 14 hours of training, the accuracy rate for the recognition of consonants showed dramatic improvement. For example, in the recognition of /r/, which was the most improved of all the consonants trained, the accuracy is reported to have increased from 36.1% to 88.6% after the training [13].

Although studies on the potential benefits of auditory-visual speech training for L2 learners has only recently started to gain focus, the importance of lip shapes as beneficial cues has long been recognized by language instructors in teaching English as a second language. For example, a study claiming that the degree of difficulty that lies in acquiring the phonemes of a foreign language may be due to the difference in visual cues had already been published more than 40 years ago [14]. In this study, the difficulty that Japanese learners of English face in making the distinction between /r/ and /l/ is taken up. While this difficulty is usually attributed to the fact that these two phonemes do not exist in the language, Goto claims that the difficulty is due to the fact that there is the disadvantage of not being able to “read the lips of the speaker”.

Using computer based methods for visual speech in language learning is still a fairly new enterprise; however, several studies have attempted to test its effectiveness. The talking head Baldi, for example, was used to teach non-native phoneme contrasts. The improvement of both speech identification and production had been observed, but the results indicated that the viewing of the internal articulators was not an additional benefit [15]. In another study, native English speakers were tested on their pronunciation accuracy of non-native segments using Bao (the Mandarin speaking version of Baldi) and Badr (the Arabic version). The analysis showed support for the value of employing visual speech in learning a non-native segment, but here again, the outside of the face was more beneficial compared to the sagittal view illustrating the tongue, palate and velum [16].

Some studies, however, have shown that the information of vocal tract articulator movements can assist in pronunciation instruction. Furthermore, the output of visual articulatory 2D or 3D models are claimed to be correctly interpreted even by young children, thus implying the usefulness of these types of models [17].

4. Segmental vs. Suprasegmental

In addition to the type of information to include in the system, another aspect that needs to be taken into consideration is the phonological unit to be focused upon. The general trend nowadays is to lay emphasis on the communicative factor, and there are several studies suggesting that focusing on the suprasegmental element in pronunciation teaching has an impact on the comprehensibility of learners’ output [e.g. 18, 19].

There is agreement among some researchers that the suprasegmental errors observed in L2 speakers have more serious effect on intelligibility than segmental ones. In one study, two groups of L2 students of English received instruction in segmental and suprasegmental features respectively. The result of this study indicated that in terms of narrative reading, only pronunciation teaching based on suprasegmental features had any effect on the comprehensibility of the learners’ production [20]. But here again, we find inconsistent results in the literature. For example, Jenkins, based on data collected from six learners of English, maintained that instruction in segments should be prioritized over suprasegmentals. The learners, two Japanese, three Swiss-German, and one Swiss-French, were instructed to engage in various pair work that included social conversation, information exchange and problem solving tasks. When analyzing the interaction that took place between the receiver and the interlocutor, Jenkins found that out of the 40 cases where the receiver could not understand the intended meaning of his/her interlocutor, 27 were designated as cases of difficulty in producing segments. Based on this finding, she concluded that instruction in segments should be prioritized.

Although there is no denying the fact that both segmental and suprasegmental aspects play significant roles in pronunciation, instruction in segments may be a good starting point, especially for beginners of L2. Since pronunciation involves the physical aspect of language, it would be beneficial for learners of L2 to locate the muscles that they need to make the individual sounds of the foreign language, especially those that do not exist in their native tongue before focusing on the higher prosodic levels. In this sense, we decided that our priority in developing the pronunciation system would be on the segments.

5. The system

Based on the findings of previous studies, we decided that our pronunciation system will be characterized by the following two features: 1) making good use of visual data; 2) focusing on the segments rather than the suprasegmentals.

As we have already mentioned above, the contribution of visual cues to the understanding of individual speech sounds goes back several decades and there is nothing new about the concept itself. The explanation taken from a traditional pronunciation drill book states that “you should feel your lips and tongue move and your jaw drop lower, then rise again as
you go from one sound to the other. Use a mirror to watch your mouth produce the sounds.” [21]. However, the major problem to this “mirror” approach as well as the other “traditional” approaches based on visual cues is that the movement of the tongue cannot be made clear since it hides inside the mouth.

The advent of computers has allowed learners to display visual cues of sound in a more “sophisticated” way. Technology now allows learners to easily convert sound data into sophisticated digital representations. However, this type of representation can be quite useless because in most cases the users would not have the slightest idea of how to interpret them. Even if one gets formal training in how to interpret these representations, it is still hard to link them into actual physical action.

Some systems take a more user-friendly approach to pronunciation by presenting a colorful and easy-to-understand illustration of the speech organ together with detailed explanation of the phonemes used in different languages. The website of the University of Iowa is one such example where an animated articulatory diagram and video clips of the sound spoken in context are presented in a user-friendly way [22]

![Figure 3: The side view](image3.png)

![Figure 4: The front view](image4.png)

In addition to these two angles, there is the transparent mode as indicated in Figure 5:

![Figure 5: The transparent mode](image5.png)

The transparent mode allows users to take a look into the internal organs in actual movement. The user is able to rotate the angle at any degree that he/she likes while slowing the...
speed of animation, or taking pauses or freezing the frame in cases where they would like to replicate the movements that take place. At present, we are compiling a list of English words which are claimed to be commonly difficult for Japanese to learn, but instead of just presenting the model movements for these words, we have added on controllers that allow instructors to freely move and adjust the lips, tongue, jaws so that they can point out the deviant movements made by the learners themselves.

At present, the lip, tongue and jaw movements have to be adjusted manually by the instructors, but we are planning to link the visual representations to a speech recognition system for use in autonomous learning.

In addition to learners of foreign languages, this system is also designed to contribute in training the speakers of pathological speech. The speech organs in transparent mode is expected to be especially useful for the hearing impaired population as well as others with severe aural and oral communication difficulties.

6. Conclusion

In this paper, we reported on the pronunciation system that we are developing. The two main features of the system are: making good use of visual data and focusing on the segments rather than the suprasegmentals. This is a practical and productive pronunciation software that should be enjoyable to use for foreign language learners of all ages.

Although it is not easy to prove the effectiveness of employing the internal articulatory movements for training, it may be possible to compare and monitor the articulatory changes made before and after the training using ultrasound images [23]. This we leave for future research.

7. Acknowledgements

The author would like to thank Fumiki Teratani, Tomohide Kano, Masahito K. Tachibana and Akira Ishii for their contributions to this study. This study is funded by Grants in Aid for Scientific Research (C) (No. 23520594).

8. References