Information Engineering Express
International Institute of Applied Informatics
2015, Vol. 1, No. 1, 1 – 9

The Utility of the SQLite Database for An Utterance Training Method Based On Lip Movements

Yuko Hoshino [*], Eiki Wakamatsu [†],
Mitsuho Yamada [‡]

Abstract

We previously proposed a utterance recognition system that uses the tracking and analysis of speakers' lip movements as words and sentences are uttered. In this report, we compared the lip movements in the SQLite database with those of utterances produced by students learning the method. We also built an utterance database from a group of text files of the utterances of Japanese TV announcers and native English speakers. We tested our system using the relational database management system SQLite for this training method.

Keywords: utterance, training method, lip movements, database, SQLite, faceAPI.

1 Introduction

The number and variety of physical settings in which the recognition of spoken information (i.e., voice recognition) is necessary are increasing, and the development of personal computers (PCs), tablets and smartphones etc. that use voice recognition is ongoing. A fundamental challenge in this research and development is to create utterance recognition that works in settings that are noisy or have many people speaking at the same time. Another obstacle to utterance recognition is settings in which it is difficult to speak loudly/clearly. In our research, we hypothesized that the use of image reorganization may be used to help voice recognition systems; that is, if a system can distinguish what the user is saying by recognizing the movements of the user's lips, the system could be used even in noisy settings and in settings in which speaking loudly is not possible. Such a recognition system may also be beneficial for individuals with hearing impairment.

Another application could be for practicing sounds in a new language, as the user can be taught the lip movements for the new language. This technique is already being used to help Japanese students learn English [1]. A 1998 report by Hayamizu et al. [2] described voice recognition technology that uses multiple modes including sound and images, but only one image was

[*] School of Information and Telecommunication Engineering Tokai University, Tokyo, Japan
[†] Graduate School of Information and Telecommunication Engineering Tokai University, Tokyo, Japan
[‡] School of Information and Telecommunication Engineering Tokai University, Tokyo, Japan

mentioned. Nagata et al. investigated the use of two microphones with directivity in an attempt to reduce the effects of ambient noise [3]. These two systems are not easily adopted, however.

An attractive and more easily used method was proposed in 2008 by Yanagi and Yamada (an author of the present study) [4]; it involves multimodal recognition based on the recognition of lip movements. The lips are first detected from the image, the motion of the lips is then analyzed, and finally the sound or word that was spoken is recognized. As shown in Figure 1, the method uses five points placed on the mouth and chin of the speaker. The movements of these five markers when the speaker utters vowel sounds, words, phrases and sentences are tracked in chronological order, creating a movement history of the coordinates. In addition, the power spectrum of each point of the lip and mouth movement is submitted to Fourier transform. The correlations between the power spectra and the spectra of an existing dictionary are then obtained, enabling the recognition of words and sentences based on the correlations.

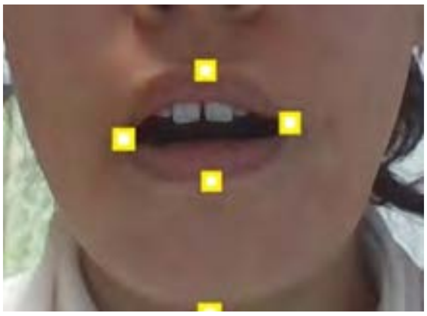


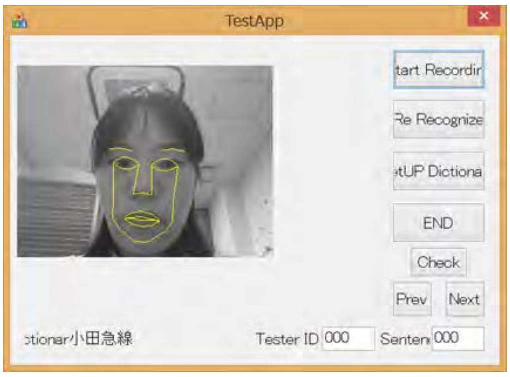
Figure 1: Lip and chin marker points

This method has been applied for lip-reading education for deaf children [5], but the imitation of the lip and mouth movements is not easily accomplished, especially for beginners. We recently improved the system so that it can compare the learner's data on a display with the correct utterance movements [1]. The dictionary is a compilation of CSV (comma-separated values) data files of the correct movement data, and the user's data file can be added and compared in this particular application's system.

2 The Creation Of The Dictionary Database

The dictionary for the utterance training method described above was created using utterances in both Japanese (by native Japanese-speaking television announcers) and native English-speaking teachers of English. Each lip movement dictionary was made from image data recorded by a digital video recording system with a built-in PC, at 640*480 (VGA) resolution. The software program faceAPI (Seeing Machines Corp., Canberra, Australia) is incorporated in the recording application software that we developed for the recognition of the position of the speaker's face, eyes and mouth as the speaker utters words and sentences.

When the distance between the speaker's outer eye corners is at least 40 pixels (according to the faceAPI specifications), the accuracy of the faceAPI software is 1 cm or less. In our video shoot conditions, the distance between the speaker's outer eye corners is four or more times this 40-pixel minimum, and we thus contend that the accuracy of the system's tracking of the movements of lips is sufficient. In addition, the processing speed of the system is 0.3 sec when the speaker's head does not move and a 2.4-GHz Intel Core-2 Duo processor is used. Our system uses a 1.8-GHz Intel i7 processor with 8 GB of memory, which provides sufficient processing speed for the detection of lip movements. This recording application is included in our new utterance training method system. A sample data acquisition screen is shown in Figure 2.

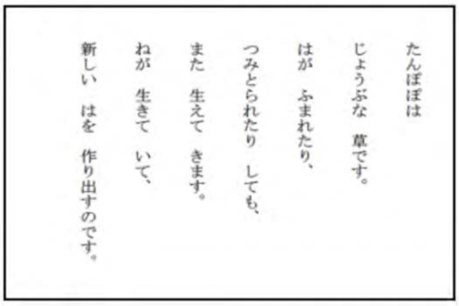


ctionar小田急線

Tester ID 000 Senten 000

Figure 2: The display menu at the time of data acquisition

The application software was used to record data for a dictionary of Japanese words, phrases and sentences. The speakers were two television announcers (one male, one female) from the NHK (Japan Broadcasting Corp.) Communication Training Institute. We used the application software to record sentences from two Japanese textbooks: "[Easy training for a good voice]" [6] and "Tanpopo [dandelion]" [7], a primary school textbook. An example Japanese sentence is provided in Figure 3. We also added many Japanese words that are frequently used in daily life, such as the names of subway stations.



たんぽぽは
じょうぶな　草です。
はが　ふまれたり、
つみとられたり　しても、
また　生えて　きます。
ねが　生きて　いて、
新しい　はを　作り出すのです。

Figure 3: A sample sentence used for Japanese utterance training

English-language dictionary data were also collected using the application software. The speakers were one male and one female English teachers who were native speakers of English. We recorded the speakers uttering 12 sentences consisting of 65 phrases. These sentences are included in an NHK language module, "Three-Month Topic English Conversation," for learners of English [8].

Our application software, which we developed previously, maintains all of these dictionary data and the learner data in CSV text files, which do not require special software to read or write. The dictionary database and software program are thus easy to install and use with other PCs. The management of the CSV text files is not difficult, since the volume of words and sentences is not high.

3 Utterance Recognition Results With Our Application System

Using the application system described above, we compared the lip movement history of a learner and the ability of the dictionary database to recognize an uttered word. For the first experiment, we used some station names of the Odakyu train line in Japan as the dictionary data. In the experiment, the learner uses the recording application with faceAPI and the dictionary data, uttering a word (in this case, a station name) toward the web camera with the built-in PC. The application outputted the movement of the five points of the learner's lips and mouth to a file, and then the power spectrum of these data from the learner was obtained for the left corner of the mouth as a sample. A Fourier transformation was conducted with the power spectrum, and the power spectrum could then be outputted as a file. The same process was done for the other marker points (i.e., left, right, top, bottom), all of which go to a file. The learner's spectra were then compared with the dictionary's spectra. Figure 4 shows the structure of the dictionary's spectra for the word some station names in the dictionary named "Odakyu."

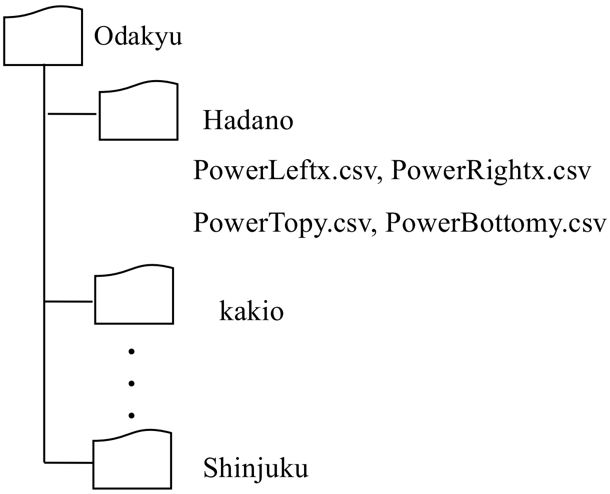

Odakyu
Hadano
PowerLeftx.csv, PowerRightx.csv
PowerTopy.csv, PowerBottomy.csv
kakio
Shinjuku

Figure 4: Structure of the dictionary "Odakyu" version

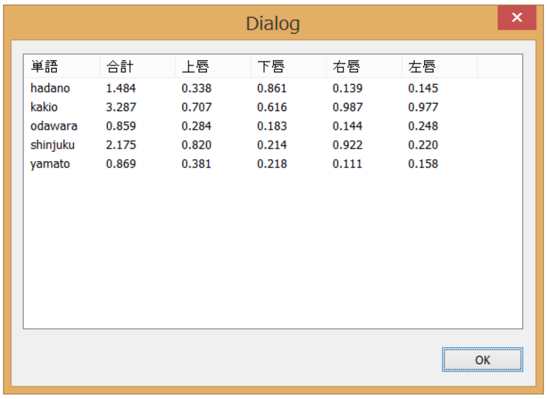
Additions to the dictionary require the power spectral data for each word, which go to a sub-folder under the dictionary name designated by the user; the spectrum file is copied there.

When a learner uses this recognition system, the protocol is as follows. When the learner pushes the 'check' button to call up the utterance recognition function, the left-mouth spectrum data are read by the system. The left-mouth spectrum data of the first word in the dictionary is then read. Next, the correlation between the power spectra of the learner and that of the dictionary are then calculated and outputted. The same is done for the learner's right, top and bottom markers, providing the complete comparison of the learner's utterance and the dictionary database using CSV. The processing is repeated for all of the words uttered, in accord with the algorithm shown in Figure 5.

When the correlation obtained for one of the markers in near 1.00, the utterance is accepted as similar to the dictionary's data, and when the sum of the four correlations is near 4.00 (for all four markers), the learner's lip movements are considered similar to the dictionary values. An example of a practice result is given in Figure 6. Based on this result and similar outcomes, we found that the reorganization of utterances from lip movements is possible at the level of the single words dictionary ("Odakyu").

```
For n: 1..number of words in the dictionary
    For each m : left, right, top, bottom
        Read learner's power$_m$ spectrum
        Read power$_m$ spectrum of word$_n$
              in Dictionary
        Calculate the correlation$_{mn}$ of learner's
              power$_m$ spectrum and power$_m$ spectrum
              of word$_n$ in Dictionary
        Output the correlation$_{mn}$
```

Figure 5: Algorithm used to calculate the correlations between the learner's utterances and the dictionary database



単語	合計	上唇	下唇	右唇	左唇
hadano	1.484	0.338	0.861	0.139	0.145
kakio	3.287	0.707	0.616	0.987	0.977
odawara	0.859	0.284	0.183	0.144	0.248
shinjuku	2.175	0.820	0.214	0.922	0.220
yamato	0.869	0.381	0.218	0.111	0.158

Figure 6: Example of a practice result for the word "Kakio")

4 Problems using csv data

4.1 Present conditions using CSV files

We created a folder for every dictionary and several subfolders in each dictionary folder for the words enrolled. The group of CSV files made from each dictionary utterance is saved in a subfolder. A folder is made for every learner, too, and his or her utterance data are saved in the subfolder as a CSV file group.

The dictionary data are increased by adding CSV files. A potential problem with this protocol is when a user puts the data in the wrong destination, especially as the number of words and number of dictionaries increase. Users may not notice if they overwrite a file with the wrong word's data. Another problem occurs when a word is not recognized. This may be solved by adding ID numbers to each dictionary and word when they are saved, with numeric data about the lip movement. As the number of files increases, the number of times that files are input or output also increases, which could cause a delay in processing.

4.2 Accumulation and analysis of learner data

The existing utterance training application stores the latest lip movement history, with which the application can reproduce the lip movement by line drawing on the system's screen display. As illustrated in Figure 7, by comparing the lip movement of the dictionary with the lip movement of the learner, the system can provide advice to the learner such as "put the lower lip higher." However, this advice is provided only for the latest utterance. If the learner repeats an utterance continuously, the system provides confirmation of the learner's progress by using the learner's prior utterance data, which are stored in CSV files. The training data are not only used for this purpose; the data can also be analyzed and used to help provide specific training feedback, such as "weak in the pronunciation of 'A'," "weak in a certain field," and "tendency to lower the lower lip."

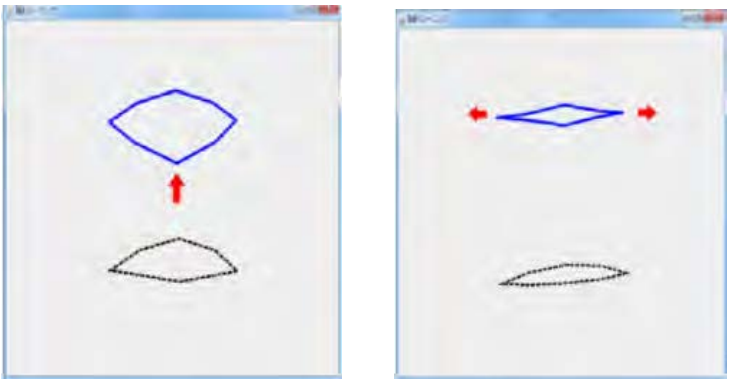


Figure 7: An example of the practical use of the utterance training, for the utterance of "ga" (left) and "I" (right)

The data for a specific purpose (e.g., the type of word or vowel sound) must be selected from among the stored data, and the processing becomes complicated when using CSV files. This problem remains to be solved.

4.3 Application to tablets and smartphones

Our system currently works on Windows as the platform, and we are investigating its use with smartphones and tablets (e.g., for Android and iOS). Smartphones usually have both their main memory and external memory (such as an SD card) for media storage. Depending on the external memory of the smartphone, authorization for media storage may be necessary; an opening sentence may thus not be possible when the developer puts the destination in external memory or an internal storage device in a program. In addition, as noted above, the number of files to be written increases each time a dictionary is updated, e.g., when a new word is registered.

5 Use Of The Database Management System

We inspected a characteristic and the performance about Relational Database and the document-oriented database with SNS data [9][10]. Those works and in light of the issues described above, we examined the use of a database management system (DBMS).

5.1 The choice of DBMS

Each dictionary is given a dictionary name, such as 'Odakyu' or the name of a vowel sound, and the numeric data for the utterance spectra are also included. Since these data require accuracy and consistency, a relational database management system (RDBMS) is desirable. When this method is used with a tablet or smartphone, a very large DBMS cannot be used, because these devices have a smaller memory capacity compared to a PC. With the database on a server, the inconvenience of poor transmission quantity increases with every use, and portable devices cannot be used when they cannot connect with the server. It would be best if the recognition method could be used without network communication. Therefore, a DBMS that can be operated in smartphones and tablets without special software is needed. As few files as possible is also desirable for the updating of the dictionary files. Based on these reasons, we chose the public-domain database SQLite (www.sqlite.org)[11].

5.2 The SQLite Database

The SQLite database is a type of RDBMS which uses one database as one file, similar to Microsoft's Access DBMS. Developers can thus use SQLite as a function call of the API library with applications. SQLite can be adopted to a standard library (e.g., that of the Android OS), and it can also be used with iOS and Windows. Moreover, the SQLite database uses storage that does not depend on the byte order, and therefore the stored files are available in the differing application software programs among the different operating systems and architectures, although there are different versions of storage files. The transaction speed of SQLite is comparable to those of the most commonly used databases, and it can be used in a variety of applications and can be operated in the native code.

5.3 Incorporating SQLite

We incorporated SQLite in an utterance training application for Windows as follows. First, the dynamic link library of SQLite and a header file of the computer language C were downloaded. Next, we added call DLL and the operation schema to the database as a substitute for the file input and output into the program. The DB files are made using the general SQL sentence (i.e., 'Create Database,' 'Create Table') by a program or a command-line tool. We compared the processing times of the database with those of the CSV files. When the quantity of data was not too large, the processing times did not show major differences.

We next compared the quantity of coding. When CSV files were used, it was necessary to use every data file to create a folder path and to process open files. An example program for a CSV file is shown in Figure 8.

With the use of the SQLite database, the dictionary data needed to be connected only one time. Moreover, a dictionary could be updated to only one file. An example program for all of the dictionary data in the SQLite database is given in Figure 9.

```
((CEdit*)GetDlgItem(IDC_EDIT1))->GetWindowText(FolderPath);
paths = (const char*)(FolderPath + "\\full2.csv");
fp = _tfopen( paths, _T("r"));
```

Figure 8: Sample program for one CSV file

```
sqlite3_open(“dic.db”, $dicDB);
```

Figure 9: Sample program for all dictionary data in the SQLite

6 Conclusion

We here introduced an utterance recognition method that uses lip movement, and we described its application to utterance exercises. We noted the problems that accompany the maintenance of the dictionary and learner data in the application software, and we incorporated the database SQLite as a means to address these problems. We found that further development in the data management is necessary. Using the data accumulated to date in the SQLite database, we are going to gather more feedback from learners/users in order to improve the method's efficiency.

A face detection library is incorporated in Android platforms from 4.0 onward. This can be used to track the position of the entire face, the center of the right eye, the center of the left eye, and the center of the mouth. The iOS5 mobile operating system has a similar function. In light of this function/library, we plan to investigate whether our utterance recognition method using lip movement with faceAPI software is possible with only a smartphone.

Acknowledgement

We thank Mr. Atsuyuki Takahashi, the NHK announcers and the English teachers who cooperated in the creation of the utterance database. This work was supported by JSPS KAKENHI Grant No. 25330418.

References

[1] Wakamatsu E, Hoshino Y, Yamada M. Proposal for an utterance training method based on lip movements. IMQA 2014 (The Seventh International Workshop on Image Media Quality and its Applications), September 2014, Chiba, Japan. pp. 44-47.

[2] Hayamizu S, Takezawa T. Trends in research on multimodal information integration system. Trans Jpn Soc Artif Intell. 2006:13(2):206-211, Mar.1998. (in Japanese).

[3] Nagata Y, Fujioka T, Abe M. Target signal detection system using two directional microphones. IEICE Trans A. 2000:J83-A(12):1445-1454. (in Japanese).

[4] Yanagi T, Yamada M. Proposal and verification of lip movement model about utterance recognition interface without voice. HIS2008 (8th International Conference on Hybrid Intelligent Systems), Sept. 2008, p. 2422. (in Japanese)

[5] Oda M, Ichinose S, Oda S. Development of a pronunciation practice CAI system based on lip reading techniques for deaf children. Technical Report of IEICE. 2007:107(179):53-58. (in Japanese)

[6] Fukushima M. Easy training for a good voice. Seibido Publishing, Tokyo. 2006. (in Japanese).

[7] Hirayama K. Dandelion. First Volume of Revised New National Language Two, Tokyo Shoseki, Jan. 1985. (in Japanese)

[8] Three-month topic English-conversation. NHK publication, 2009. (in Japanese)

[9] Hoshino Y, Consideration about the Database Storage to Manage Twitter Data, IMQA 2013 (The Sixth International Workshop on Image Media Quality and its Applications), September 2013, Tokyo, Japan. pp. 166-169.

[10] Hoshino Y, Fujino I, Comparison between SQL and NoSQL using Twitter data, The 8th National Convention of Japan Personal Computer Application Technology Society, December 2013, Osaka, Japan, C2-3. (in Japanese)

[11] public-domain SQLite Web Site, http://www.sqlite.org/