INTRODUCTION TO THE rtMRIDB TRIAL VERSION

Kikuo Maekawa
(National Institute for Japanese Language and Linguistics)

1. What is rtMRIDB?

The real-time MRI articulatory movements database (rtMRIDB) is a collection of movies that captured articulatory movements of the vocal tract producing various speech sounds. The mid-sagittal plane image of a vocal tract is captured by a specially operated MRI device at the rate of about 14 fps. The database systematically visualizes the dynamic aspects of articulatory phonetics to which enough attention has not been paid so far. The current "trial-release" version contains about 13,000 records collected from ten speakers of Standard or Tokyo Japanese. The public release of the final version is expected in a few years from now on.

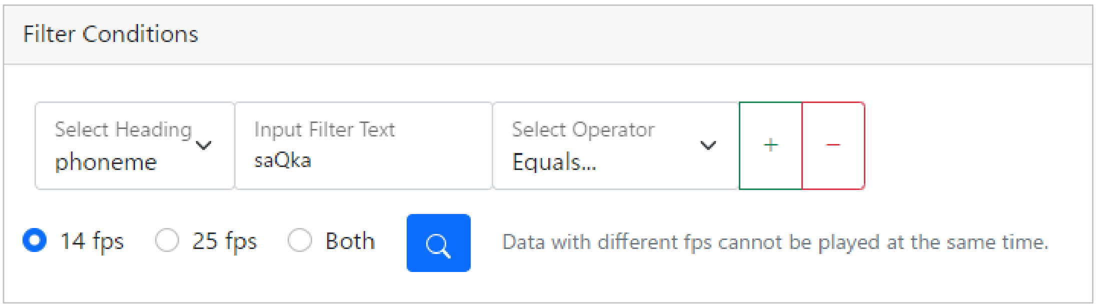
2. How to use the database

2.1. Simple query

To make a search of this database, users need to specify the "search conditions" on the web site. The way you choose necessary records is very similar to the use of the filter function in MS Excel. Figure 1 shows an example. Select "phoneme" in the "Select Heading" column, specify "saQka" in the "Filter Text" field, and select "Equals …" in the "Select Operator" field. Then click on the magnifying glass icon to start the search.

In this case, there are 11 hit, and the metadata of the 11 records are displayed on the screen as in Figure 2. You may wonder why there are 11 hits even though there are only 10 speakers. This is because, in this case, the subject s7 has uttered this item twice for some reason. On the left edge of each record in Figure 2, there is a check box allowing you to select/deselect the record. By default, all the hit records are selected (marked with ✓), hence they are all included in the output movie.

Figure 1: Example specification of search condition



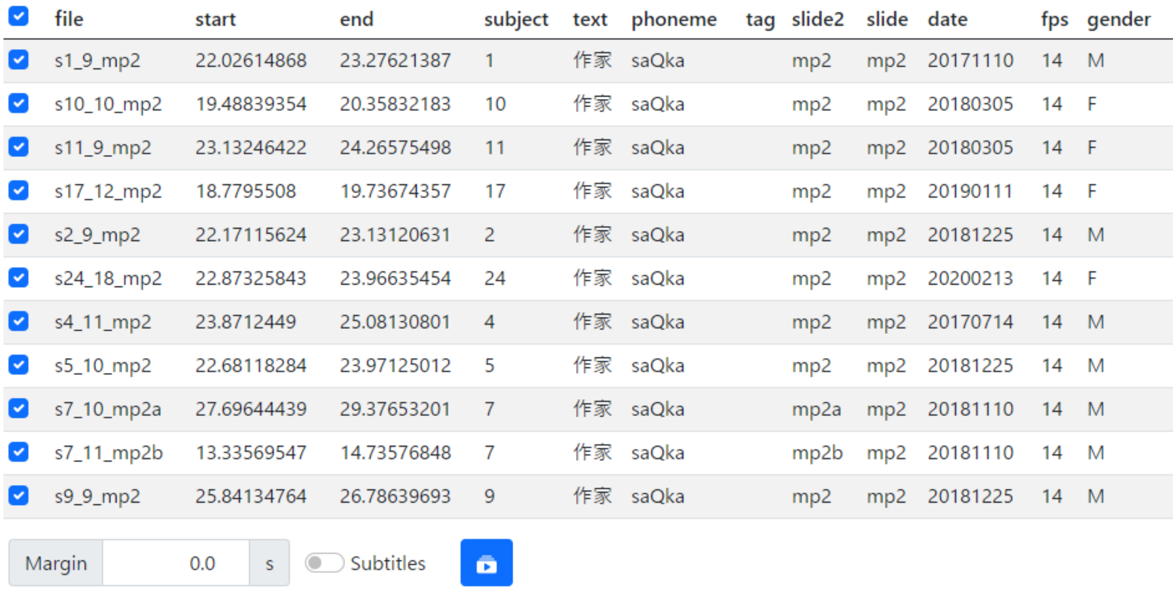
Filter Conditions

Select Heading: phoneme
Input Filter Text: saQka
Select Operator: Equals...

● 14 fps ○ 25 fps ○ Both

Data with different fps cannot be played at the same time.

Figure 2: Result of the search (Some of the right-most columns are not shown)



☑	file	start	end	subject	text	phoneme	tag	slide2	slide	date	fps	gender
☑	s1_9_mp2	22.02614868	23.27621387	1	作家	saQka		mp2	mp2	20171110	14	M
☑	s10_10_mp2	19.48839354	20.35832183	10	作家	saQka		mp2	mp2	20180305	14	F
☑	s11_9_mp2	23.13246422	24.26575498	11	作家	saQka		mp2	mp2	20180305	14	F
☑	s17_12_mp2	18.7795508	19.73674357	17	作家	saQka		mp2	mp2	20190111	14	F
☑	s2_9_mp2	22.17115624	23.13120631	2	作家	saQka		mp2	mp2	20181225	14	M
☑	s24_18_mp2	22.87325843	23.96635454	24	作家	saQka		mp2	mp2	20200213	14	F
☑	s4_11_mp2	23.8712449	25.08130801	4	作家	saQka		mp2	mp2	20170714	14	M
☑	s5_10_mp2	22.68118284	23.97125012	5	作家	saQka		mp2	mp2	20181225	14	M
☑	s7_10_mp2a	27.69644439	29.37653201	7	作家	saQka		mp2a	mp2	20181110	14	M
☑	s7_11_mp2b	13.33569547	14.73576848	7	作家	saQka		mp2b	mp2	20181110	14	M
☑	s9_9_mp2	25.84134764	26.78639693	9	作家	saQka		mp2	mp2	20181225	14	M

Margin 0.0 s ○ Subtitles ▶

The selected records will be concatenated into a single video file and are played back if you click on the ▶ icon at the bottom of Figure 2. The video format is MP4. The pop-up video playback window can be closed by clicking the "Close" button at the bottom right of the window.

If you turn on the "Subtitle" switch on the left side of the ▶ icon before playing back the video, the file name and the start and end times will be displayed as subtitles in the lower right corner of the video, as in Figure 4. The integer following the "s" in the beginning of the file name (e.g. "s1_9_mp2" in Figure 4) corresponds to the speaker's ID. You can use the "Margin" box to the left of the "Subtitle" switch if you want to play a wider range than the start and end times specified in the metadata. Note that if you specify too large a value, the end of the previous utterance or the beginning of the following utterance (or both!) will be played back. Normally, 0.0 works well.

Figure 3: Example of a window playing back the concatenated video file

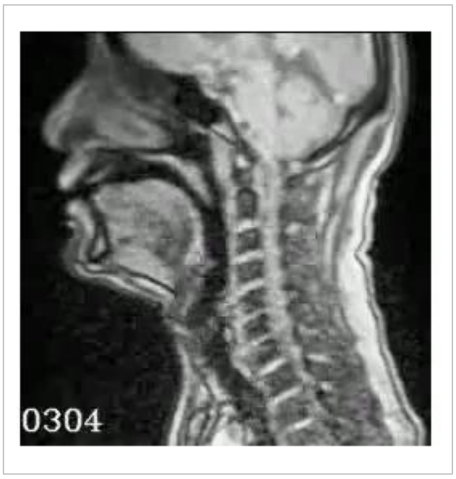
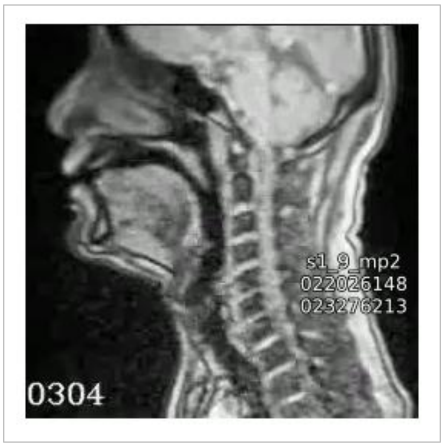


Figure 4: Example of a video with subtitle



2.2 Notes on search technique

2.2.1 Difference from Excel

There are a few issues that you should be aware of when searching this database. The first thing to note is that, unlike the Excel's filter function, the pull-down menu of "Input Filter Text" does not show a list of items contained in the field. The user needs to enter the text by referring to the text or phoneme column in Appendix. See Section 9 below about the appendix. Another important difference from the Excel is that our database is case sensitive, i.e. the uppercase and lowercase letters are distinguished as different characters.

2.2.2 Text field

Some of the items in the "text" field contain information other than words in a narrow sense. For example, if you specify "辛酸" (phonemically /siNsaN/ 'hardship')as the text and "equals …" as the operator, you will not get any hits. You need to specify "辛酸（しんさん）" instead, because, in the metadata, this word is registered not by itself but followed by an annotation on the pronunciation shown within the parentheses. In doing so, you will have 11 hits. The same results will be obtained, alternatively, by specifying "辛酸" as the text, and "begins with" as the operator. You can find in the appendix the inventory of the strings recorded in the text field of the database.

2.2.3. Phoneme field

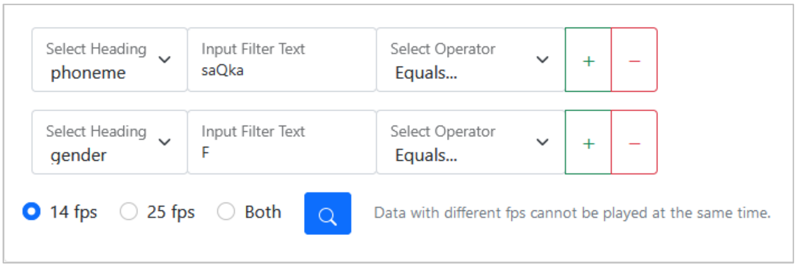
Similar problems occur when specifying a text in the phoneme field. Some of the strings in the phoneme field contain subscripts (like _1 or _2) to distinguish homonyms, and some others contain hyphens to indicate differences in word structure. For example, the text corresponding to the phoneme string "eHgo_1" is "英語" ('English')and the text corresponding to phoneme string "eHgo_2" is "A5" (size of paper). Also, the phoneme string "saka_1" corresponds to the text "坂" ('slope'), but the unsubscripted “saka” corresponds to an mb item "サカ", which is a nonsense sequence of two morae. Similarly, the phoneme string for "砂糖屋" ('suger vendor') is "satoH-ja", and the phoneme string for "里親" ('foster parent') is "sato-oja". In this case, the hyphen indicates a morpheme boundary, but please note that not all morpheme boundaries are indicated, only those that are necessary for disambiguation are indicated. Please check the appendix to see the inventory of the phoneme strings.

2.3. Specification of multiple search conditions

2.3.1 Specifying multiple conditions for different fields

It is possible to specify multiple conditions for searching. In this case, the specified multiple conditions are connected, in principle, by the AND (i.e., logical product) operator(s). For example, in Fig. 5, the condition that "gender" is equal to "F" is additionally specified to the search condition shown in Figure 1. To add a new condition, click on the + sign icon displayed on the right side of the condition. A window for specifying the new condition will open. If you run the search with the condition shown in Figure 5, only four records of female speakers will be hit. To delete a part of the conditions after specifying multiple conditions, click on the - sign icon on the right side of the condition you want to delete.

Figure 5: Example specification of multiple search conditions



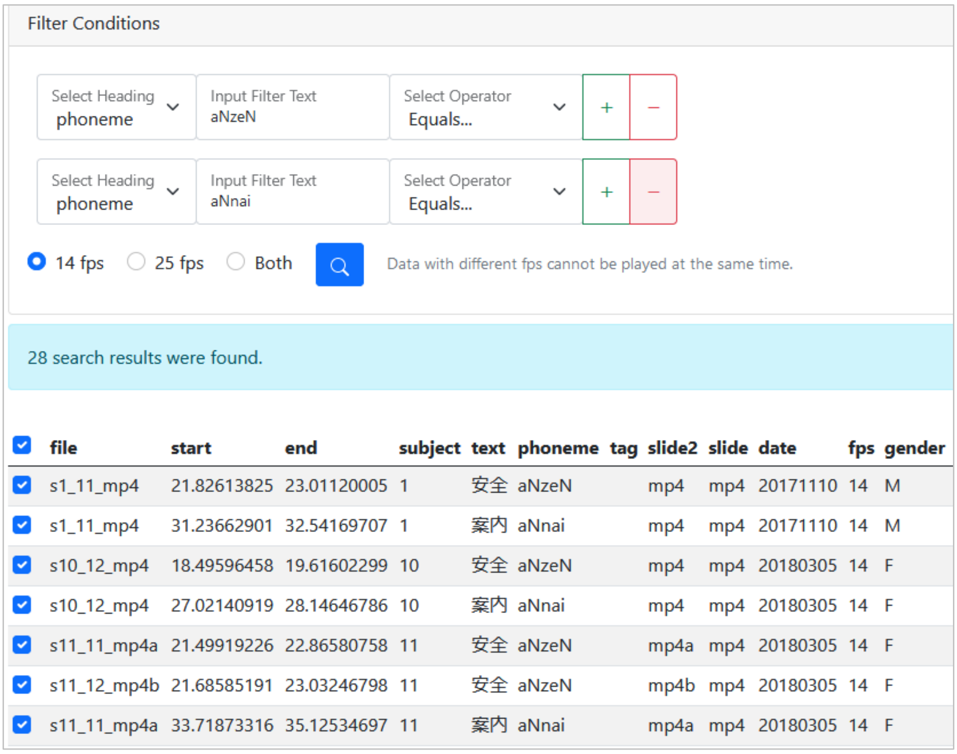
Select Heading: phoneme | Input Filter Text: saQka | Select Operator: Equals...

Select Heading: gender | Input Filter Text: F | Select Operator: Equals...

● 14 fps ○ 25 fps ○ Both

Data with different fps cannot be played at the same time.

Figure 6:　Example specification of multiple search conditions for the same field



Filter Conditions

Select Heading: phoneme | Input Filter Text: aNzeN | Select Operator: Equals...

Select Heading: phoneme | Input Filter Text: aNnai | Select Operator: Equals...

● 14 fps ○ 25 fps ○ Both

Data with different fps cannot be played at the same time.

28 search results were found.

☑	file	start	end	subject	text	phoneme	tag	slide2	slide	date	fps	gender
☑	s1_11_mp4	21.82613825	23.01120005	1	安全	aNzeN		mp4	mp4	20171110	14	M
☑	s1_11_mp4	31.23662901	32.54169707	1	案内	aNnai		mp4	mp4	20171110	14	M
☑	s10_12_mp4	18.49596458	19.61602299	10	安全	aNzeN		mp4	mp4	20180305	14	F
☑	s10_12_mp4	27.02140919	28.14646786	10	案内	aNnai		mp4	mp4	20180305	14	F
☑	s11_11_mp4a	21.49919226	22.86580758	11	安全	aNzeN		mp4a	mp4	20180305	14	F
☑	s11_12_mp4b	21.68585191	23.03246798	11	安全	aNzeN		mp4b	mp4	20180305	14	F
☑	s11_11_mp4a	33.71873316	35.12534697	11	案内	aNnai		mp4a	mp4	20180305	14	F

2.3.2 Specifying multiple conditions for the same field

In the above example, multiple search conditions were specified for different fields (phoneme and gender), but it is also possible to specify multiple conditions for a single field. Note, however, that in such a case, the specified conditions will be connected by the OR operator(s). In Figure 6, "aNzeN" ('safety') and "aNnai" ('introduction') are specified for the phoneme field (using "equals to" operator). The search results contain 28 records including both "aNzeN" and "aNnai".

2.4. Database fields that can be utilized for database search

In addition to the "text" and "phoneme" fields, you can also specify search conditions for the "subject", "tag", "slide2", "slide", "date", "fps", "gender", "birthYear", and "dialect". A brief description of these fields is given in Table 1 (a more detailed explanation is given in Section 9). The fields in Table 1 are chosen in view of the final public release version. Currently, however, some of them are not yet useful for the search. For example, all fps are "14", and all dialects are "Standard", so they are not useful for narrowing down the list. Tag field contains a note on each record, but the density of the tag information in the current version differs considerably depending on when it was assigned.

Table 1: Fields that can be utilized in the database search

Field	Content
subject	Speaker ID（Currently ten subjects）
text	Japanese text shown in the utterance list. Sometimes annotated with respect to the pronunciation.
phoneme	Pseudo-phonemic representation of the text
tag	Annotation on pronunciation and articulation (currently inconsistent across speakers)
slide2	Subscript like a, b, c, etc. showing the repetition of the same slide. No subscript when there is no repetition.
slide	ID given to each slide consisting the utterance list
date	Date of recording
fps	Frame rate of the rtMRI movie　（currently 14 exclusively）
gender	Gender (sex) of the subject
birthYear	Birth year of the subject
dialect	Dialect spoken by the subject（Curretnly Tokyo Japanese）

3. Speakers

Table 2 shows the properties of the ten speakers recorded in the trial-release version. They consist of six male and four female speakers of Standard Japanese, and are identified by a corresponding speaker ID that begins with a letter 's' followed by speaker-specific integers, like 's1' and 's2'. All speakers are college educated. Speakers, s1, s2, s4, s5, s7, and s9 are researchers in the field of phonetics or speech engineering.

Table 2: Properties of the ten speakers

ID	**Gender**	**BirthYear**	**BirthPlace**	**RecordingDate**
1	M	1956	Kyoto	20171110
2	M	1970	Tokyo	20181225
4	M	1969	Tokyo	201707
5	M	1958	Tokyo	20181225
7	M	1955	Tokyo	20181110
9	M	1990	Kanagawa	20181225
10	F	1967	Saitama	20180305
11	F	1971	Tokyo	20180305
17	F	1969	Tokyo	20190111
24	F	1956	Tokyo	20200213

4. Utterance list

The data contained in the current trial-release version were collected during the years 2017-2020. The utterance list used in the data collection has been expanded during the course of the data collection. As a consequence, some items in the list were not uttered by all ten speakers. For example, the item /kadaN/ ('flower bed') in slide mp5 was uttered only by three speakers of s4, s17, and s24 (See appendix). Currently, supplementary recording is on the way so that we can make the final version of the database more balanced.

The utterance list consists of several utterance classes, but only the data of three classes, classes mu, mp, and mb, are included in the current database. The contents of these classes will be explained in the following subsections.

In the recording of an rtMRI data with the rate of 14 fps, it was possible to record the utterance of up to 37 seconds long. Speakers were requested to read aloud the items printed on a slide (shown in a screen in the MRI machine) with in the time limit. There are times, however, when speakers couldn't read aloud all items in a slide within the allotted time. It happens, typically, when the speaker repeated some items due to mispronunciation. When it happened, we repeated the recording of the same slide. As a consequence, most of the items in the slides were recorded correctly more than two times. This is why some items in the current database are uttered multiple times by the same speakers. The multiple records of the same slide uttered by the same speaker are distinguished by an alphabetic subscript like 'a', 'b', or 'c' given to the end of the slide name. When a slide was read only once (which is the most frequent case), there is no subscript in the slide name.

4.1. Mora unigram (the mu class)

Items in the mu class are the recording of articulatory movements behind each mora of the Standard Japanese produced without career sentence. The utterance list of the mu class is similar to the "50 mora table" taught in the elementary schools in Japan, but differs from the pedagogical table in that the items in the mu class covers much wider area of possible morae in the Standard Japanese. As of December 2020, 110 morae are included in the mu class; and 32 sequences of two morae (like /ai/, /au/, /ira/, /iri/ etc.) are also included in the mu class. Exactly speaking, the two-morae sequences are not mora unigram, but they are included in the mu class for the sake of convenience. Some items in the mu class are not uttered by all ten subjects, because the list has expanded during the course of the recording (See appendix).

4.2. Mora bigram (the mb class)

Items in the mb class consist of 676 combinations of the 26 morae /ka, ki, ku, ke, ko, kja, kju, kjo, sa, si, su, se, so, sja, sju, sjo, ma, mi, mu, me, mo, ha, hi, hu, he, ho/. These bimorae like /kake/, /sami/, or /hakja/ are produced using the career sentence of /korega ___ gata/ ('This is type ___'). This career sentence has two characteristics: for one, it controls the lexical accent of the target bimora by the accent-removing suffix of /gata/; in addition, the same /ga/ morae are located on both sides of the target bimorae. The number of items in the mb class is constant regardless of the date of recording, but on some rare occasions, some items are lacking due to mispronunciation (See appendix).

4.3. Mora phoneme (the mp class)

Items in the mp class are collected to investigate the so-called mora phonemes (or special morae) of Japanese, i.e., the moraic nasal /N/, geminate /Q/, and long vowel /H/. The diphthong /J/ is usually counted as a mora phoneme, but in this database the samples of diphthong were not collected systematically, though there are pairs like /sjoHkai-zjoH/ ('letter of introduction') versus /sjoHka-izjoH/ ('digestive disease') that can be distinguished in terms of the difference between the diphthong and two sequential vowels. The number of recorded items in the mp class differs considerably from one speaker to another depending on the date of recording. See the appendix for the details.

5. Making of the MP4 videos

The MRI images of the mid-sagittal profile of vocal tract were collected using the MRI and related equipment settled in the ATR Brain Activity Imaging Center (ATR-BAIC) in Kyoto (MAGNETOM Prisma fit 3T, Siemens). With this machine, it is possible to make recording of up to 37 seconds of articulatory movements with the frame rate of about 14 fps; the result being 512 frames of static MRI images each consisting of 256 x 256 pixels with the slice width of 10 mm. The MRI data is recorded in the standard DCM format. At the same time, the speech signal was recorded using a DAT with 44.1 kHz and 16-bit sampling. The MP4 movies contained in the current database was converted from the original DCM data. The speech recorded by the DAT was also dubbed to the MP4 file. The time alignment between the DCM images and DAT signal was made by making reference to the timing where the MRI machine started to make its operational noise. Note there is possibility that the time alignment is not perfect due to the large difference in the sampling rates of the images and speech signal. Note also that weak noise reduction is applied to the audio signal in the MP4 file to reduce the operation noise of the MRI machine.

6. Acknowledgement

We owe much to Mr. Takuya Asai of Waseda University (The Graduate School of Human Science) for the construction of MP4 movies. The prototype version of the rtMRIDB was designed by Mr. Ken'ya Nishikawa of the National Institute for Japanese Language and Linguistics. The database construction was supported by the JSPS KAKENHI grants (17H02339, 19K21641, 20H01265) and a special budget from the President of the National Institutes for the Humanity. We thank Drs. Shinobu Masaki, Yasuhiro Shimada, and Yukiko Nota of ATR-BAIC for their support in the data acquisition.

7. Reference

For detailed information about the rtMRIDB project and the MRI techniques used in the rtMRI data acquisition see the document below.

K. Maekawa, K. Nishikawa, T. Asai, Y. Nota, S. Masaki, Y. Shimada, N. Takemoto, T. Kitamura, Y. Saito, T. Kagomiya, Y. Ishimoto, H. Kikuchi, M. Fujimoto, and Y. Yagi. "Design of Real-Time MRI Articulatory Movement Database." *Proc. Language Resource Workshop 2020*. Center for Corpus Development, The National Institute for Japanse Language and Linguistics, September 2020. [In Japanese] https://pj.ninjal.ac.jp/corpus_center/LRW2020PDF/P3-5.pdf

The following paper provides an example how rtMRIDB contributes to the study of phonetics.

Kikuo Maekawa. "Production of the utterance-final moraic nasal in Japanese: A real-time MRI study". *Journal of the International Phonetic Association*, in press for open access.

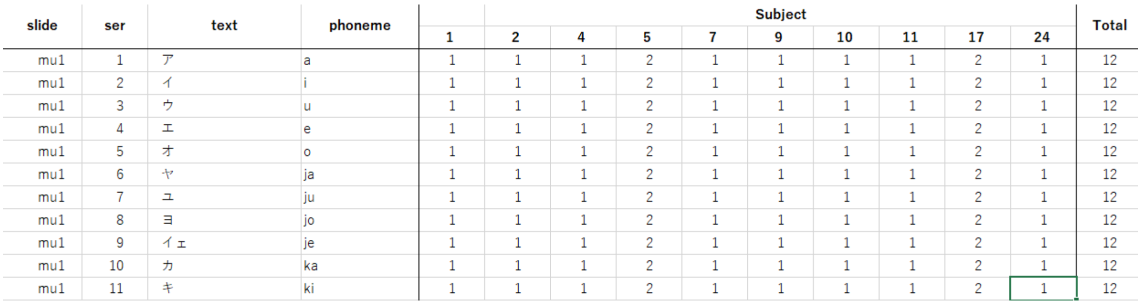
Limitations of this database

The final version of this database is planned to be released under CC BY, but since many problems are expected to exist in this trial -release version, CC BY will not be applied to the database. However, you are free to make link to this database but please don't forget to let us know. You are also free to cite some of the images in this database in your academic writings on condition that you are aware that this database will be modified in the near future. Reuse or redistribution of this database is not permitted until it is fixed as the final version.

Contact address: kikuo*ninjal.ac.jp (Please replace the “*” symbol with an at sign.)

9. Appendix: Relationship between speech items and speakers.

The appendix is prepared to indicate how many times each of the 969 items in this database was spoken by each speaker. Click here to read the appendix.



slide	ser	text	phoneme	Subject										Total
				1	2	4	5	7	9	10	11	17	24	
mu1	1	ア	a	1	1	1	2	1	1	1	1	2	1	12
mu1	2	イ	i	1	1	1	2	1	1	1	1	2	1	12
mu1	3	ウ	u	1	1	1	2	1	1	1	1	2	1	12
mu1	4	エ	e	1	1	1	2	1	1	1	1	2	1	12
mu1	5	オ	o	1	1	1	2	1	1	1	1	2	1	12
mu1	6	ヤ	ja	1	1	1	2	1	1	1	1	2	1	12
mu1	7	ユ	ju	1	1	1	2	1	1	1	1	2	1	12
mu1	8	ヨ	jo	1	1	1	2	1	1	1	1	2	1	12
mu1	9	イェ	je	1	1	1	2	1	1	1	1	2	1	12
mu1	10	カ	ka	1	1	1	2	1	1	1	1	2	1	12
mu1	11	キ	ki	1	1	1	2	1	1	1	1	2	1	12

(The beginning of the appendix)

"Slide" is the ID of the slide presented to the speaker during the experiment, where the first two letters indicate the speech class (mu, mp, mb). The first two letters indicate the utterance class (mu, mp, mb); Ser is the order of speech within a slide; for example, in the case of the mu item, there were 27 or 32 speech items on a slide; Text is a Japanese text indicating how each speech item was printed on the slide; in some of the mp items, the pronunciations were annotated in parentheses.