Web application to convert English into helpful characters for pronunciation learners

Keita Nakatsuka[1,*], *Akitsugu* Nogita[1,**], and *Ian* Wilson[1,***]

[1]CLR Phonetics Lab, University of Aizu, Tsuruga, Ikki-machi, Aizuwakamatsu, Fukushima-ken, Japan

Abstract. The International Phonetic Alphabet (IPA) is the most commonly used set of phonetic symbols but it can be difficult to understand and too abstract for non-phoneticians, such as English learners and foreign language educators. One of the factors that makes it difficult for English learners is the number of vowels used in English. In previous research, this problem was solved by proposing a condensed list of 13 English vowels and 24 consonants that are logically organized for North American English, and by assigning a new phonetic symbol font called *Sound Spelling* to these phonemes. However, there are currently no English texts that have *Sound Spelling* to show learners this easy way to pronounce and it is difficult for non-phoneticians to write transcriptions right away, because materials using those symbols are lacking for English learners. In this research, we developed a web application, now publicly available, that converts input English into *Sound Spelling* quickly and accurately to solve this problem.

1 Introduction

Our research is the development of an application that helps to improve the pronunciation of English learners by using a new phonetic symbol system called *Sound Spelling*, which is used instead of the International Phonetic Alphabet (IPA) [1]. In previous research, Ueda and Otsuka [2] showed that there are substantial differences in how different companies (dictionary publishers, etc.) handle pronunciation guidance items, and it is pointed out that the textbook used has a great influence on the learner. Nogita [3] showed that the list of phonetic symbols in IPA notation in English-Japanese dictionaries and pronunciation materials, especially the list of vowel symbols, contains a surplus and also dialect differences, resulting in a number of symbols far more than the actual number of phonemes in one dialect. Since it is difficult for an English language learner to grasp the whole picture of pronunciation, Nogita proposed a condensed list of 13 English vowels and 24 consonants that are logically organized for North American English – especially Canadian English. Also, a phonetic symbol font system called *Sound Spelling*, which is based on English speaking Children's phonics and spelling books, was assigned to these phonemes. This *Sound Spelling* is useful because non-phoneticians only have to learn a 13-character sound set to talk with accurate pronunciation that native speakers can understand when focusing only on vowels. However, there are currently no English texts that have *Sound Spelling* to show learners this easy way to pronounce and to give examples to teachers for creating practice materials. It is difficult for non-phoneticians to write transcriptions right away in a timely manner, because materials using those symbols are lacking for English learners. Our research is to develop a publicly available web application that converts input English into *Sound Spelling* quickly and accurately to solve this problem. Eventually, we aim to incorporate *Sound Spelling* into actual English teaching and learning through the web application.

2 Some phonetic symbol sets for English

2.1 International Phonetic Alphabet

The International Phonetic Alphabet (IPA) is the most commonly used set of phonetic symbols. However, it can be difficult to understand and too abstract for non-phoneticians, such as English learners and educators. It is especially tough for people who are learning English for the first time. One of the factors that makes it difficult is the number of vowels used in English. For example, the list of phonetic symbols in IPA notation in English-Japanese dictionaries and some pronunciation materials in authorized textbooks, especially the list of vowel symbols, contains a great surplus and is subject to dialect differences, making the total number of symbols far more than the actual number of phonemes in one dialect.

2.2 ARPAbet

ARPAbet [4] is a phonetic symbol set developed by Advanced Research Projects Agency (ARPA) as a part of their Speech Understanding Research project in the 1970s.

---
*e-mail: s1240213@u-aizu.ac.jp
**e-mail: anogita@u-aizu.ac.jp
***e-mail: wilson@u-aizu.ac.jp

It represents phonemes and allophones of General American English with distinct sequences of ASCII characters. ARPAbet has been used in several speech synthesizers, and it is also used in the Carnegie Mellon University (CMU) Pronouncing Dictionary [5].

2.3 Sound Spelling

Sound Spelling is a phonetic symbol set developed by Nogita [3] by organizing pronunciation-based spellings used in English speaking children's phonics and spelling books for the purpose of establishing instruction to grasp the overall picture of the English pronunciation system. It consists of a condensed set of 13 English vowels and 24 consonants that are logically organized for North American (especially Canadian) English. This *Sound Spelling* set is useful because non-phoneticians only have to learn a 13-character vowel set to talk with accurate pronunciation that native listeners can understand when focusing only on vowels.

The correspondence between each of these three phonetic notations (i.e., IPA, ARPAbet, and *Sound Spelling*) can be seen in Table 1 for 13 vowels and 24 consonants.

3 Method

To achieve the goals of this research, a database had to be created first. The database has all English words in *Sound Spelling* form. Then a webpage could be created that calls the database when the user inputs English text.

3.1 Database creation

3.1.1 CMU pronouncing dictionary

To create a database, we first needed a dictionary with one-to-one correspondence between English words and pronunciation. In this research, we downloaded the dictionary *cmudict-0.7b.txt*, distributed by Carnegie Mellon University (CMU) under the BSD license, from the CMU webpage [5] and used it as the source of the database. The dictionary has more than 134,000 items of data and covers all common words. *Sound Spelling* is a phonetic symbol set logically organized for North American English, and the CMU dictionary is also made for North American English. The dictionary is well managed and expanded. For these reasons, we considered it appropriate to use this dictionary. A sample of the format of the data from file *cmudict-0.7b.txt* can be seen in Figure 1.

3.1.2 Conversion from ARPAbet to Sound Spelling

Since *cmudict-0.7b.txt* has phonetic symbols in ARPAbet, *Sound Spelling* is not immediately available. So we must convert the phonetic symbols to be expressed in *Sound Spelling*. However, since *cmudict-0.7b.txt* has more than 134,000 entries, it would be very tedious to change every single word in the database. Instead, we created a PHP program that instantly converts ARPAbet notation into *Sound Spelling* notation.

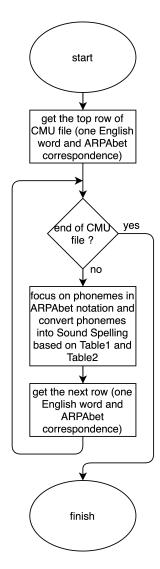**Table 1.** Correspondence between IPA, ARPAbet and *Sound Spelling* for 13 vowels and 24 consonants in various sample words

13 Vowels	IPA	ARPAbet	*Sound Spelling*
hate	eɪ	EY	ā
Pete	i	IY	ē
site	aɪ	AY	ī
note	oʊ	OW	ō
cute	u	UW	o͞o
hat	æ	AE	ă
pet	ɛ|e	EH	ĕ
sit	ɪ	IH	ĭ
not	ɔ|ɑ	AO or AA	ŏ
cut	ʌ|ə	AH or AX	ŭ
coin	ɔɪ	OY	oy
loud	aʊ	AW	ow
book	ʊ	UH	o͝o

24 Consonants	IPA	ARPAbet	*Sound Spelling*
buy	b	B	b
pie	p	P	p
die	d	D	d
tie	t	T	t
vie	v	V	v
fight	f	F	f
guy	g	G	g
kite	k	K	k
high	h	HH or H	h
joy	dʒ	JH	j
China	tʃ	CH	ch
lie	l	L	l
my	m	M	m
nigh	n	N	n
rye/turn	ɹ/ɻ	R	r/ur
zoo	z	Z	z
sigh	s	S	s
wise	w	W	w
yacht	j	Y	y
pleasure	ʒ	ZH	zh
shy	ʃ	SH	sh
they	ð	DH	dh
thigh	θ	TH	th
sing	ŋ	NX or NG	ng

```
ABOARD  AH0 B AO1 R D
ABODE  AH0 B OW1 D
ABOHALIMA  AE0 B AH0 HH AH0 L IY1 M AH0
ABOLISH  AH0 B AA1 L IH0 SH
ABOLISHED  AH0 B AA1 L IH0 SH T
ABOLISHES  AH0 B AA1 L IH0 SH IH0 Z
ABOLISHING  AH0 B AA1 L IH0 SH IH0 NG
ABOLITION  AE2 B AH0 L IH1 SH AH0 N
ABOLITIONISM  AE2 B AH0 L IH1 SH AH0 N IH2| Z AH0 M
ABOLITIONIST  AE2 B AH0 L IH1 SH AH0 N AH0 S T
ABOLITIONISTS  AE2 B AH0 L IH1 SH AH0 N AH0 S T S
```

Figure 1. Sample format of CMU's Pronouncing Dictionary file *cmudict-0.7b.txt*

In creating the PHP program, a helpful program was an MIT-licensed program [6] to convert *cmudict-0.7b.txt* from ARPAbet to IPA. We improved it and created a new program to convert *cmudict-0.7b.txt* from ARPAbet to *Sound Spelling*. A flowchart of the program is shown in Figure 2.

Figure 2. Flowchart of the PHP program that converts ARPAbet to *Sound Spelling*

First, we will examine the *cmudict-0.7b.txt* lines from top to bottom. There is an English word and ARPAbet in one line, and the English word and ARPAbet are separated by two single-byte spaces. Next, we pay attention to ARPAbet and make sure that the ARPAbet is not empty. Then, we split the ARPAbet into phonemes. We look at the phonemes one by one. If phonemes exist, they are converted into *Sound Spelling* according to correspondence Table 1. When all ARPAbet phonemes have been converted into *Sound Spelling*, we write the English words and *Sound Spelling* into a new .txt file. We repeat this sequence until the last line of *cmudict-0.7b.txt*.

3.1.3 Addition of syllabification and syllable stress functions

Being able to convert ARPAbet to *Sound Spelling* is helpful, but it does not show syllable and stress information that are important for pronunciation. Therefore, a program was created to add this information.

The first step in adding syllable and stress information is to add syllable information. This is because, according to phonetics, one syllable is made up of consonant(s) + vowel + consonant(s) (CVC), and the stress varies from syllable to syllable. Any consonants before the CVC vowel are called the *onset*, and any consonants after the vowel are called the *coda* of the syllable. There are various ways to divide English pronunciation into syllables using a program, but this time we divided it into syllables according to the complex branching coda rules shown in one phonetics textbook [7]. They are rules stating that there are about 60 different onsets and they consist of at most three consonants. A flowchart for that is shown in Figure 3.

First, the English word in the *Sound Spelling* notation is divided on the right side of each vowel, and each segment is stored in an array. Next, we check whether the last segment contains a vowel. If the last segment does not contain a vowel, then that segment is regarded as the last syllable coda and it does not continue to the loop. Then, from the second segment from the beginning of the array, we determine if each segment has one onset out of about 60. If so, we add a hyphen to the left of each onset to separate them into syllables using a loop. Adding stress information involves simply adding into each of the syllables a primary stress symbol with upper case letters and optional secondary stress information using only upper case letters (with no symbol).

For example, the English word *captain* is *KAE1PTAX0N* in ARPAbet notation in the CMU Pronouncing Dictionary. The letters in that ARPAbet notation are converted to kă1ptŭn in the *Sound Spelling* notation (with the number 0 removed). Then, that word is divided on the right side of each vowel to create kă1 ptŭ n, and each segment is stored in an array. Next, we check whether the last segment contains a vowel. The last segment *n* does not contain a vowel, so that segment is regarded as the last syllable's coda and is attached to the second last segment, giving kă1 ptŭn. Then, from the second segment from the beginning of the array ptŭn, we determine if each segment has one onset out of about 60 allowable ones. The array ptŭn has one onset *t* out of the 60, so we add a hyphen to the left of onset *t* to separate it into syllables kă1p-tŭn. Finally, the primary stress level indicated in ARPAbet (by the number *1*) is shown by converting that syllable to upper case letters with a stress symbol, giving the final ˈKĂP-tŭn. For words with a secondary stress level indicated by the number *2* in ARPAbet, upper case letters are used with no stress symbol.
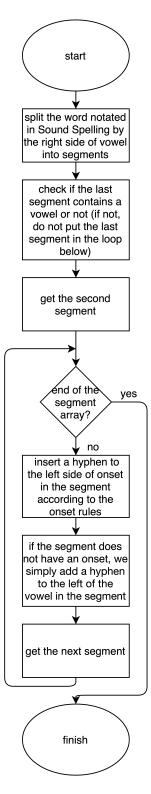
Figure 3. Flowchart of the syllabification program

3.1.4 Registration of .txt file into MySQL database

The next step in the database creation was to register the text file into a MySQL database. To do that, various preparations were made first. We started MySQL and executed the following with MySQL statements: First, we created one database and one table. Next, we created four fields in the table. The *id* field stores the primary key. The *en-*

glish field stores English words. The *ss* field stores *Sound Spelling* corresponding to English words. The *modified* field stores the date and time when the data was registered.

After those preparations, we imported the created .txt file into the table using the following MySQL statement:

LOAD DATA INFILE '~ /filename.txt' INTO TABLE 'table name' fields terminated by '\t' (english,ss);

3.2 Webpage creation

The webpage that was created in this research is made up of multiple files: HTML, PHP (ver. 7.3.7), CSS (ver. 3), and Javascript. First, the HTML file is used to explain the purpose of the research and about *Sound Spelling*. We prepared HTML files written in English and Japanese to support users accessing the website in either language. Next, the PHP file is used to add the ability to convert English to *Sound Spelling* on the webpage. We also prepared PHP files written in English and Japanese. With these PHP files, the webpage can access the database, which contains about 134,000 pairs of English words and *Sound Spelling*. In a PHP file, one can additionally write by HTML. So, when creating the page that converts English to *Sound Spelling*, we use one PHP file to access the database with the description of PHP and obtain the necessary data, and display on the screen with the description of additional HTML. Then, CSS files and Javascript files are used to determine the design of the webpage. Bootstrap 3 is also used. To maintain consistency, the webpage design was made in the same style as that of the Center for Language Research Phonetics Laboratory at the University of Aizu [8].

4 Resultant webpage

The resultant webpage is shown in Figure 4 and it is publicly available at http://clrlab1.u-aizu.ac.jp/convert_e.php. It shows results that were converted to *Sound Spelling* for all vowels and consonants of General American English using typical words (e.g., off, father, bee, ... yes, way) User-inserted English text is moved to the bottom of the screen and *Sound Spelling* corresponding to that is then displayed in the center [9].

5 Challenges and future work

A functioning website to convert English into *Sound Spelling* was created. However, various challenges were encountered. When syllable stress was to be expressed, the primary stress was planned to be in bold and upper case letters, and the secondary stress was to be simply in upper case letters, but it was difficult to express bold characters from PHP. As before, the primary stress symbol (') is tentatively attached. A second challenge is that there are words with different pronunciations and meanings, such as "read" and "increase", even with the same spelling. In future work on this webpage, it would be helpful to solve these problems.

Figure 4. Screenshot of the actual working webpage [9]

References

[1] G.K. Pullum, W.A. Ladusaw, *Phonetic Symbol Guide* (The University of Chicago Press, 1996)

[2] H. Ueda, T. Otsuka, Bulletin of Osaka Jogakuin University **7**, 15 (2011)

[3] A. Nogita, Studies on Foreign Languages and Cultures XXVIII (Kokushikan University) **28**, 30 (2018)

[4] J.E. Shoup, *"Phonological Aspects of Speech Recognition" in Trends in Speech Recognition* (Englewood Cliffs: Prentice Hall, 1980), pp. 125–138

[5] Carnegie Mellon University, *The CMU Pronouncing Dictionary*, http://www.speech.cs.cmu.edu/cgi-bin/cmudict

[6] W.E. dos Santos, *Arpabet-to-IPA*, https://github.com/wwesantos/arpabet-to-ipa

[7] Y. Hattori, *Phonetics* (Asakura Publisher, 2012)

[8] *The CLR Phonetics Lab*, http://clrlab1.u-aizu.ac.jp

[9] *Website to convert English into Sound Spelling*, http://clrlab1.u-aizu.ac.jp/soundspelling_e.html