Pertanika J. Sci. & Technol. 31 (3): 1291 - 1311 (2023)

PERTANIKA JOURNALS

SCIENCE & TECHNOLOGY

Journal homepage: http://www.pertanika.upm.edu.my/

Review Article

A Review on Grapheme-to-Phoneme Modelling Techniques to Transcribe Pronunciation Variants for Under-Resourced Language

Emmaryna Irie, Sarah Samson Juan* and Suhaila Saee

Faculty of Computer Science and Information Technology, Universiti Malaysia Sarawak, 94300 UNIMAS, Kota Samarahan, Sarawak, Malaysia

ABSTRACT

A pronunciation dictionary (PD) is one of the components in an Automatic Speech Recognition (ASR) system, a system that is used to convert speech to text. The dictionary consists of word-phoneme pairs that map sound units to phonetic units for modelling and predictions. Research has shown that words can be transcribed to phoneme sequences using grapheme-to-phoneme (G2P) models, which could expedite building PDs. The G2P models can be developed by training seed PD data using statistical approaches requiring large amounts of data. Consequently, building PD for under-resourced languages is a great challenge due to poor grapheme and phoneme systems in these languages. Moreover, some PDs must include pronunciation variants, including regional accents that native speakers practice. For example, recent work on a pronunciation dictionary for an ASR in Iban, an under-resourced language from Malaysia, was built through a bootstrapping G2P method. However, the current Iban pronunciation dictionary has yet to include pronunciation variants that the Ibans practice. Researchers have done recent studies on Iban pronunciation variants, but no computational methods for generating the variants are available yet. Thus, this paper reviews G2P algorithms and processes we would use to develop pronunciation variants automatically. Specifically, we discuss data-driven techniques such as CRF, JSM, and JMM. These methods were used to build PDs for Thai, Arabic, Tunisian, and Swiss-German languages. Moreover, this paper also highlights the importance of pronunciation variants and how they can affect ASR performance.

Keywords: Automatic speech recognition, G2P technique, grapheme-to-phoneme, pronunciation variants, under-resourced language

ARTICLE INFO

Article history:
Received: 31 May 2022
Accepted: 04 November 2022
Published: 31 March 2023

DOI: https://doi.org/10.47836/pjst.31.3.10

E-mail addresses:
emmaryna.ji@gmail.com (Emmaryna Irie)
sjsflora@unimas.my (Sarah Samson Juan)
ssuhaila@unimas.my (Suhaila Saee)
*Corresponding author

ISSN: 0128-7680
e-ISSN: 2231-8526

INTRODUCTION

Iban language is an isolect or neutral language of the Austronesian language family (Sutlive, 1994). The iban language is associated with the Malayic subgroup, the Malay language, but they are not the same. The difference gap is big in terms of spelling, pronunciation, and meaning. The addition rule of prefixes and suffixes still applies to the Iban language.

According to Sutlive (1994), there were initially 22 phonemes in the Iban language. Currently, 34 phonemes of the Iban language are found in the studies made by Juan et al. (2015). The recent study focused on the Iban language in general, noting the words and phonemes of the standard Iban language. These phonemes have potential pronunciation variants, especially in the Iban language. The Iban language has no slang or dialect (Sutlive, 1994). Nevertheless, variations of pronunciation in the language exist, distinct from each other throughout the Sarawak region, as stated by Shin (2021). The region or area is the river from which it came, for example, the Engkari river, Undop river, Sebuyau river, and Baloh River.

The Iban language is considered an under-resourced language in terms of language technology and application. Under-resource languages are mostly less studied, and lack digitalised documents because there are few language resources, and languages are passed down verbally through generations (Singh, 2008). Other examples of under-resourced languages from different parts of the globe are Tunisia (Masmoudi et al., 2016), Bangla, or Bengali (Chowdhury et al., 2018), and Thai language (Rugchatjaroen et al., 2019; Saychum et al., 2016). For Iban, a language corpus exists with thirty-one thousand (31k) Iban words collected and used for building an Iban ASR (Juan et al., 2015). Language corpora here refers to the collection of written or spoken texts that can be used to analyse speech patterns. The written texts will be laid out or listed as a pronunciation dictionary.

There are no systematic methods for generating pronunciation variants for the Iban language. Therefore, a suitable modelling technique or method is needed. This paper aims to review ASR G2P modelling techniques available for generating pronunciation variants.

BACKGROUND OF STUDY

Pronunciation Dictionary for an ASR for Iban, an Under-Resourced Language

A pronunciation dictionary is a term for a list of words that consists of word and phoneme pairing. The phonemes are phonetic symbols that comply with the International Pronunciation association (IPA) standards. They can be generated via grapheme-to-phoneme (G2P) conversion using G2P techniques, which will be discussed further in the literature review.

The current Iban ASR holds 34 phonemes of the Iban language, as mentioned and studied by Juan et al. (2015). Furthermore, the study has compiled a total of 31k Iban words, resulting from the bootstrapping method, a closely related language to the Iban language,

which is the Malay language. The experiment conducted in the study uses the Deep Neural Network (DNN) system, which resulted in a 15.8%-word error rate. The compilation was then used as the pronunciation dictionary for the Iban ASR.

Figure 1 shows a sample pronunciation dictionary Juan and Flora (2015) developed for the Iban ASR.

```
penerang        p @ n @ r a NG
tadi    t a d i KK
pesisir p @ s i s i @ r
taja    t a dZ @ KK
berita  b @ r i t a
kepala  k @ p a l a KK
sepuluh s @ p u l u @ h
aum     a w u e m
lebih   l @ b i @ h
parlimen        p a r l i m @ n
ketuai  k @ t u w aj
iban    i b a n
sibu    s i b u
mujur   m u dZ u r
asal    a s a l
kuching k u tS i @ NG
```

Figure 1. Parts of words and phoneme pairing from the Iban pronunciation dictionary

From Figure 1, the left side of the column is the Iban word or graphemes. The right side of Figure 1 shows the phonemes of the Iban word in SAMPA format. SAMPA (Speech Assessment Methods Phonetic Alphabet) is a format for phonetic script in a machine. It uses 7-bit printable ASCII characters based on IPA (International Pronunciation Alphabet). For example, Table 1 shows the Iban IPA according to Omar (1981).

Table 1 shows 19 consonant phonemes and 11 vowel cluster phonemes. Examples shown in Table 1 are given to show how it is applied in Iban's common words.

Table 1

Iban vowel and consonant phonemes with examples by Omar (1981)

	Classification	Phoneme	Place of articulation	Example
Consonant	Plosive/ stop	/p/	Bilabial	/pandak/ (short), /pintu/ (door)
		/b/	Bilabial	/badas/ (good), /baruh/ (down)
		/t/	Alveolar	/tanʧaŋ/ (tie up), /tiluək/ (scoop)
		/d/	Alveolar	/dampiəh/ (nearby), /duɲa/ (world)
		/k/	Velar	/pekakas/ (tool), /kibaʔ/ (left)
		/g/	Velar	/gagit/ (excited), /gerau/ (spook)
		/ʔ/	Glottal	/mukaʔ/ (open), /ɲemaʔ/ (if)
	Nasal	/m/	Bilabial	/majaw/ (cat), /merindaŋ/ (entertain)
		/n/	Alveolar	/menoa/ (world), /mansaŋ/ (forward)
		/ɲ/	Palatal	/meɲa/ (long time ago), /ɲirap (slice)
		/ŋ/	Velar	/eŋkabaŋ/ (light red meranti), /ŋabaŋ/ (to visit)
	Affricate	/ʧ/	Palatal	/ʧelap/ (cold), /tinʧin/ (ring)
		/ʤ/	Palatal	/ʤampat/ (hurry), /ʤera/ (guilty)

Table 1 *(Continue)*

	Classification	Phoneme	Place of articulation	Example
	Fricative	/s/	Alveolar	/sampi/ (prayer), /siŋkaŋ/ (footsteps)
		/h/	Glottal	/luhus/ (straight), /ŋehembai/ (spread out)
	Trill/ rolled	/r/	Alveolar	/riŋat/ (angry), /ŋerembaŋ/ (get across)
	Lateral	/l/	Alveolar	/telai/ (whisper), /tilok/ (scoop)
	Semi-vowel	/j/	Palatal	/gaja/ (looks like), /ukuj/ (dog)
		/w/	Bilabial	/gawa/ (work), /wai/ (calling someone older than us)
Vowel	Vowel cluster	/ai/	Fronting	/makai/ (eat)
		/ui/		/ukui/
		/ia/	Backing	/maia/ (at the time)
		/ea/		/mageaŋ/ (all)
		/ua/		/muai/ (throw away)
		/oa/		/menoa/
		/iu/		/tiup/ (blow away)
		/au/		/tauka/ (or)
		/iə/	Centering	/niliək/ (glance)
		/uə/		/puən/ (the beginning)
		/oə/		/tusoək/ (to insert a thread, to suck)

Pronunciation Variants and Why It Matters

A pronunciation variant is a term for a language that has an alternate way of speaking or different intonations of the base word and has a slight difference in terms of spelling. Examples of pronunciation variations are typically known as dialect or slang of the language. Pronunciation variants happen when two or more cultures are mixed from migration to an area (Shin, 2021).

Having a basic pronunciation dictionary is beneficial for an under-resource language, but including the pronunciation, variants can be much more beneficial. For example, some other languages have their dialects and slang added to their ASR pronunciation dictionary, as mentioned in the study by Stadtschnitzer and Schmidt (2018); Masmoudi et al. (2016).

In Stadtschnitzer and Schmidt (2018), the language in focus is Swiss German, a dialectal form of the Swiss language. The people of Switzerland use this dialect in their conversations and switch to the German language when it comes to conversing with visitors so that they can understand (Stadtschnitzer & Schmidt, 2018). The dialect is highly used in Swiss broadcasts. However, there is no standardised writing on the dialect itself. When it is tested in the ASR, the system cannot detect the language- the desired outcome is that when the dialect is inserted into the system, the output comes out as standard German writing.

The study has successfully implemented the dialect into the system by training the Swiss-German model using the standard German model, thus creating its pronunciation dictionary.

Another example of dialect added to the ASR system is the Tunisian dialect, as Masmoudi et al. (2016) studied. Including the dialect enhances the Tunisian ASR system and the pronunciation dictionary of the Tunisian language, which is called TunDPDic (The Tunisian Dialect Phonetic Dictionary). The pronunciation dictionary contains the rules and exceptions of Tunisian words, base words, and dialects. This dictionary is useful for future studies of the Tunisian language as it guides the user on the rules of base and dialect words when developing an ASR system of the language.

Common languages such as English and French also have variations in their dictionary. Take English as the prime example, where two common English instances are separated by variations or accents- British English and American English. The accents have different dictionaries for each other. For instance, Accents of British Isles corpus (Tjalve & Huckvale, 2005) and BEEP dictionary (http://svr-www.eng.cam.ac.uk/comp.speech/Section1/Lexical/beep.html) are for British English, while CMUDict corpus (Yolchuyeva et al., 2019) is for American English. The Cambridge Dictionary and Oxford Dictionary have compiled variations of the English language, including pronouncing and spelling the words. The most straightforward comparison of spelling and pronunciation of the words are such:

- The word "water" is pronounced as /ˈwɔtəɹ/ (w-a-t-e-r) in American English, /ˈwɔːtə/ (w-o-t-a-h) in British English.
- The spelling of the word "colour" is/ˈkʌlər/ (American English), and "colour" is /ˈkʌlə/ (British English).

There are reasons to include these pronunciation variants, both cultural and technical. Preserving the language's culture and uniqueness helps expose the differences and the importance of the language's history. When a language such as English is known to others, it can identify the speaker's ethnicity and cultural identity, thus making people recognise the existence of the language (Guazzi et al., 1983).

Furthermore, including the variants in an ASR system helps improve- but does not necessarily do- the system's quality (Karanasou, 2013). It also helps increase the probability for the ASR system to analyse word variations and the base word of the targeted language, thus enriching the system's dictionary with various possibilities. Finally, including pronunciation variants into the system can also be used as a bootstrap for another language with almost the same base as the selected language, as used in Juan and Besacier (2013), where Malay language data was used as the seed data for the bootstrapping method.

However, the current pronunciation dictionary for Iban ASR does not consist of pronunciation variants. The current pronunciation dictionary contains the standard form of Iban words. Word variants are yet to be included as the method mentioned earlier only

focuses on developing a base Iban pronunciation dictionary. The pronunciation variant of the Iban language exists according to different areas across Sarawak. Based on Shin's (2021) findings, the pronunciation variants exist because Ibanic speakers' migration from West Kalimantan happened in the 19th century. Some highlighted Iban word variants; examples are shown in Table 2.

Table 2
The comparison between base Iban word and its variants

Base word	Variant 1 (Kapit)	Variant 2 (Sibu)
Rumah (House)	*Humeah*	*Rumeah*
Urang (People or Humans)	*Uheang*	*Ureang*
Barang (Stuff, like an object)	*Baheang*	*Bareang*

From Table 2, the variants shown are the additional or replacement of letters from the base word. The additional *'ea'* instead of /a/ and the replacement of /r/ with /h/ are obvious differences in variants found in Iban speakers from different areas in Sarawak.

Thus, this paper aims to investigate grapheme-to-phoneme methods that can generate Iban pronunciation variants to improve the current pronunciation dictionary. Moreover, a general G2P framework is described in this paper to illustrate the flow of grapheme-to-phoneme conversion and selecting candidates for pronunciation variants.

LITERATURE REVIEW

Modelling pronunciation data by leveraging statistical approaches can help generate more word-phoneme pairs for a pronunciation dictionary. It can replace the manual labour efforts by linguists to transcribe all words to phonemes and reduce human mistakes during the transcribing process. We reviewed selected Grapheme-to-phoneme (G2P) modelling for the motivation in improving our current pronunciation dictionary with pronunciation variants. Furthermore, we briefly describe under-resourced language and its G2P challenges.

Grapheme-to-Phoneme Conversion

Grapheme-to-phoneme conversion (G2P) is the task of finding the pronunciation of a word in its written form (Bisani & Ney, 2008). It was also essential in ASR and TTS (text-to-speech) systems (Yu et al., 2020). The term phoneme refers to the smallest unit of sound that makes up a complete word, while grapheme refers to a letter or a group of letters to represent the sound of the phoneme. Figure 2 shows the basic flowchart of G2P conversion.

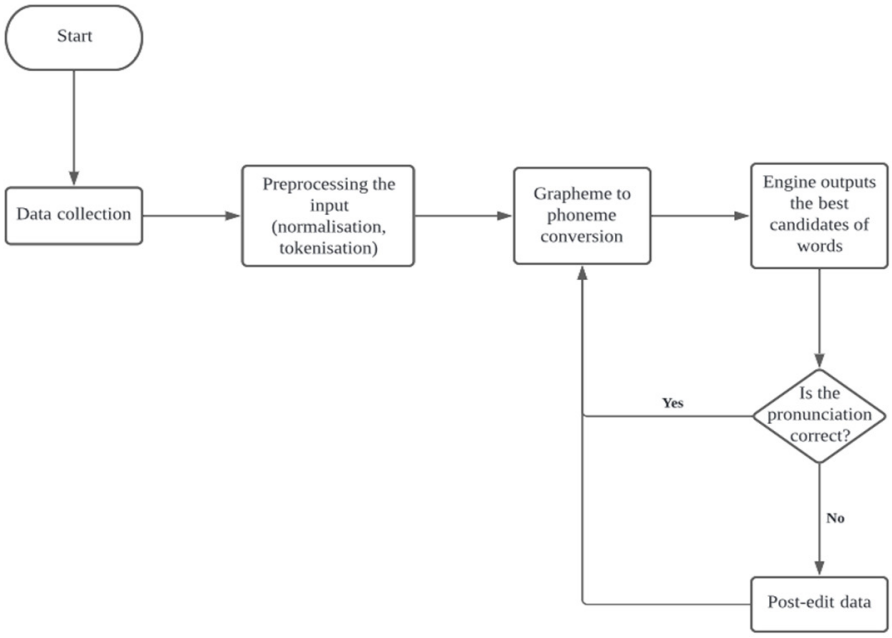

Figure 2. Basic flowchart of G2P conversion in a G2P technique

Figure 2 simplifies the data-driven flowchart from Juan and Flora (2015), which have implemented the G2P technique in their work. The following will explain the steps in detail.

Data Collection

In this process, raw text data are collected and compiled into a database or in an excel sheet. The data collection format can be formal texts, scripts, newspapers, websites, and other text-based platforms. The amount of data collected may vary from one to another, depending on the availability of the targeted language. For example, from an online news portal, Juan and Flora (2015) collected 7,000 articles in the Iban language from 2009-2012 related to sports, entertainment and general matters. From these articles, there were 2.08 million words found.

Pre-Processing the Input

The data that has been collected will be pre-processed to remove noise and unwanted symbols. This process is based on Ramli et al. (2015), where pre-processing involved tokenising the words, normalising the data into a machine-readable format, and sampling and grouping the data according to the amount of data of the chosen parameters. Lastly, the data are transformed into text in their orthographic form. An example of normalisation is as follows: the number '4' will be changed into 'four', and symbols such as *'$'* will be changed to 'dollar'. The tokenisation process includes spaces between the normalised words or letters, depending on the G2P modelling techniques. Iban data were pre-processed using

the above techniques, and the sentences were segmented to obtain 36 thousand unique words, which will then be used in the G2P conversion (Juan & Flora, 2015).

Grapheme-to-Phoneme Conversion

Converting the sequence of letters into sequences of phones is called G2P conversion. Its job is to convert a letter string into a phoneme string form (Jurafsky & Martin, 2000). The conversion process needs rules or learning from the seed data given during the process. The data-driven approach derives the pronunciation data from the seed data (Laurent et al., 2014). This study focuses on data-driven modelling techniques, which will be explained further in *Data-Driven G2P techniques*.

The G2P Engine Output the Best Candidates for Words

After the previous process, the outputs are compiled together and checked for errors. Language experts or native speakers can verify the outputs by conducting post-editing tasks, as shown in a previous study, to obtain a gold standard letter-phoneme pair (Juan & Flora, 2015). Then, the performance of the G2P model can be evaluated using metrics such as phoneme error rate, word error rate, and perplexity by comparing the gold standard with the G2P outputs (Chen et al., 1998).

Post-Edited Data

The post-edited letter-phoneme pairs can then be added to the training data to improve samples for the G2P model, and the next sequence of words is predicted using the improved model. This bootstrapping strategy reduces the time to transcribe graphemes to phonemes manually and systematically improves the quality of a G2P model as this approach can be repeated many times according to the language vocabulary size.

Previous work applied the bootstrapping strategy based on the semi-supervised method using a local dominant language, Malay, to create a base Iban phoneme sequence (Juan & Besacier, 2013). In this work, a Malay G2P was developed using an existing Malay pronunciation dictionary (Tan et al., 2009) as the source for the model. From the 36 thousand entries of the Iban word lexicon, about 1,000 words were phonetised using Malay G2P to obtain phonetic transcripts, and the outputs were post-edited to match with Iban pronunciations. Subsequently, another 1,000 words were phonetised by the same G2P, and the outputs were post-edited to get Iban phonemes. Hence, bootstrapping outputs from Malay G2P became the base for Iban G2P to convert the remaining entries in the Iban word lexicon.

Data-Driven G2P Techniques

The selected G2P modelling techniques have been used in research and experiments in the past years.

Conditional Random Fields. Conditional random fields, or CRF in short, are one of the techniques used for grapheme-to-phoneme conversion. This technique utilises a network of non-directional nodes and vertices. The nodes contain every possibility or probability of the next sequence word or letter being trained and tested. This technique shares a similar inner working with Hidden Markov Model (HMM); the only difference is that CRF did not make any assumptions about the data interdependence or independence chosen as the model (Morris, 2010). It is a framework for building probabilistic models to segment and label sequence data (Lafferty et al., 2001).

In G2P conversion, CRF defines a conditional probability distribution over label sequences by a given observation sequence rather than a joint distribution of label and observation sequences (Illina et al., 2011). Given the training grapheme or letter-to-phoneme associations and some predefined feature sets, CRF learns a set of weights *w*. The learning process of set *w* parameters is usually done by maximum likelihood learning for $p(\bar{y} \mid \bar{x};\ w)$ as in Equations 1 and 2:

$$p(\bar{y} \mid \bar{x};\ w) = \frac{1}{Z(\bar{x},\ w)} \exp \sum_j w_j F_j(\bar{x},\ \bar{y}) \tag{1}$$

$$F_j(\bar{x}, \bar{y}) = \sum_{i=1}^{n} f_j(\bar{y}_{i-1}, \bar{y}_i, \bar{x}, i) \tag{2}$$

Where,
$\bar{x}$: sequence of letters
$\bar{y}$: sequence of phonemes
w: weights
fj: feature function

The feature function can depend on the sequence of word letters, the current and previous phonemes, and the current position in the given the word. For example, in Equation 2, unigram features are shown as $f_j(\bar{y}_i, \bar{x}, i)$, while bigram features are represented as $f_j(\bar{y}_{i-1}, \bar{y}_i, \bar{x}, i)$. The unigram features will only be utilised when the bigram features use the current and previous phoneme sequence.

There has been a discussion regarding methods such as HMM, MEMM (maximum entropy Markov Model), and CRF, and ultimately, CRF can solve HMM and MEMM issues of bias labelling in experimental stages, as stated in Morris (2010) and Lafferty et al. (2001).

Joint Multigram Modelling. The joint multigram model (JMM) was pioneered by Deligne et al. (1995). It is a statistical model for matching streams of symbols under the hypothesis that all came from a common underlying stochastic process. JMM utilises the segmentation

process, in which it segments the letter of the word and phoneme and pairs them in every possible way. Figure 3 illustrates the segmentation of JMM.

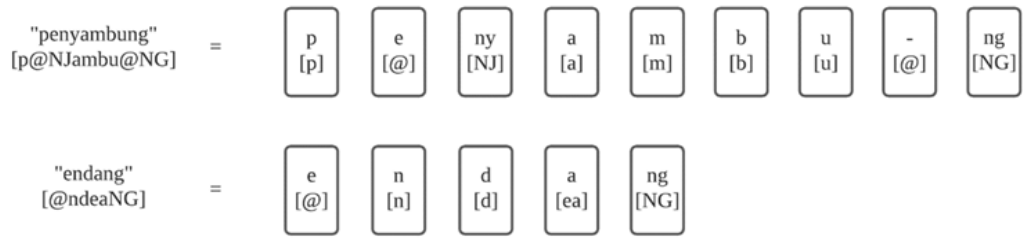

Figure 3. Examples of the segmentation process by JMM

To explain the JMM process illustrated in Figure 3, two streams of symbols in Equation 3 present a pair of sequences of $\begin{bmatrix} s_t \\ \omega_t \end{bmatrix}$.

$$\begin{pmatrix} O = o_{(1)} \dots o_\tau \\ \Upsilon = \varepsilon_{(1)} \dots \varepsilon_{(\Theta)} \end{pmatrix} \tag{3}$$

A (n, υ) JMM is a model in which the longest size of sequences in *O* and ϒ are, respectively, *n* and *υ*. Furthermore, the model allows the pair sequences s_t and ω_t to be of unequal length, which further assumes a many-to-many alignment between two strings. Taking L_o and L_Υ as a cosegmentation of *O* and *ϒ*, where *L* is the corresponding joint segmentation of *O* and ϒ, which *L* can be denoted as Equation 4:

$$L = (L_O, L_\Upsilon) \tag{4}$$

where *L* consists of all possible cosegmentation. Meanwhile, the overall likelihood of (*0*, ϒ) is computed as the sum of all cosegmentation (Equation 5).

$$L(O, \Upsilon) = \sum_{L \in \{L\}} L(O, \Upsilon L) \tag{5}$$

Assuming the concatenated consequences independent, the likelihood will be denoted as Equation 6.

$$L(O, \Upsilon, L) = \prod_t p \begin{bmatrix} s_t \\ \omega_t \end{bmatrix} \tag{6}$$

A decision-oriented version of the model approximates the likelihood of the corpus as Equation 7:

$$L^*(O, \Upsilon) = max_{L \in \{L\}} L(O, \Upsilon, L) \tag{7}$$

and defines the most likely cosegmentation of *L** as Equation 8.

$$L^* = argmax_{L \in \{L\}} L(O, \Upsilon, L) \tag{8}$$

Taking another example from an Iban word "urang" and "ureaNG", where O = *"urang"* and Υ = *"ureaNG"*; the best segmentation obtained using Equation 8 will be pictured as the following Equation 9:

$$\begin{aligned} & p\begin{bmatrix} u \\ u \end{bmatrix} \cdot p\begin{bmatrix} rang \\ reaNG \end{bmatrix},\ p\begin{bmatrix} ur \\ ur \end{bmatrix} \cdot p\begin{bmatrix} ang \\ eaNG \end{bmatrix}, \dots \\ & p\begin{bmatrix} u \\ u \end{bmatrix} \cdot p\begin{bmatrix} r \\ r \end{bmatrix} \cdot p\begin{bmatrix} ang \\ eaNG \end{bmatrix}, p\begin{bmatrix} ura \\ urea \end{bmatrix} \cdot p\begin{bmatrix} ng \\ NG \end{bmatrix}, \dots \\ & p\begin{bmatrix} u \\ u \end{bmatrix} \cdot p\begin{bmatrix} r \\ r \end{bmatrix} \cdot p\begin{bmatrix} a \\ ea \end{bmatrix} \cdot p\begin{bmatrix} ng \\ NG \end{bmatrix}, \dots \end{aligned} \tag{9}$$

All consequence probabilities are estimated on a training corpus. Besides that, the JMM model can automatically decode a test input string O into an output string Υ through a sequence-by-sequence transcription process (Deligne et al., 1995). It can be a standard maximum of a posteriori decoding problem, which consists of finding the most likely string $\hat{\Upsilon}$ given the stream of O (Equation 10).

$$\hat{\Upsilon} = argmax_{\Upsilon}\, L(\Upsilon|O) = argmax_{\Upsilon}\, L(O,\Upsilon) \tag{10}$$

Assuming $L^* = (L_O^*, L_\Upsilon^*)$ - most likely joint segmentation of O and Υ representing most of the likelihood, Equation 7 will be maximised as Equation 11:

$$\widehat{\Upsilon^*} = argmax_{\Upsilon}\, L(O, \Upsilon, L_O^*,\ L_\Upsilon^*) \tag{11}$$

By using the Bayes rule, Equation 11 can be rewritten as Equation 12:

$$\widehat{\Upsilon^*} = argmax_{\Upsilon}\, L(O, L_O^*|\Upsilon, L_\Upsilon^*)\, L(\Upsilon, L_\Upsilon^*) \tag{12}$$

where $L(O, L_O^*|\Upsilon, L_\Upsilon^*)$ measures the likelihood of the matching between O and Υ along with their best cosegmentation.

Several studies, such as Cherifi and Guerti (2021), Masmoudi et al. (2016), and Wang and Sim (2013), have been using this G2P method in their experimentations.

Joint Sequence Modelling. Joint sequence modelling, or JSM in short, was founded by Bisani and Ney (2008). The overview of JSM is that the relation of input and output sequences was generated from the common sequence of the joint unit that carried input and output symbols. The term *consequence* and *joint multigram* refers to the unit that carries multiple input-output symbols (Deligne et al., 1995; Bisani & Ney, 2008). In JSM, the joint units, the grapheme-phoneme joint multigram, were stated as graphones. Figure 4 shows an example of JSM segmentation.

An orthographic form is given a sequence of letters or characters. It is sometimes referred to as graphemes. Pronunciation, on the other hand, is represented in phonemic transcription, a sequence of phoneme symbols. By denoting a set of graphemes as ***G*** and

a set of phonemes as ϕ, the task of G2P conversion is following Bayes' decision rule to obtain the optimal phone sequence (Equation 13)

$$\varphi(g) = arg \max_{\boldsymbol{\varphi' \epsilon \phi^*}} P(\boldsymbol{g}, \boldsymbol{\varphi'}) \tag{13}$$

where in each orthographic form $g \epsilon G^*$, we seek the most likely pronunciation $\varphi \epsilon \phi^*$.

A graphone is a pair of letter and phoneme sequences of possibly different lengths. It is denoted as Equation 14:

$$q = (g, \varphi) \epsilon Q \subseteq G^* \times \phi^* \tag{14}$$

The expressions of $\boldsymbol{g_q}$ and $\boldsymbol{\varphi_q}$ referred to the first and the second component of $\boldsymbol{q}$, respectively. A graphone is mentioned as *singular* if it has one letter and one phoneme at most. In JMM, a common sequence of graphones assumes the orthographic sequence of the word and the phoneme. The letter and phoneme sequences are grouped according to an equal amount of segmentation called *cosegmentation* (Deligne et al., 1995). This segmentation uses many-to-many alignment, which has the advantage of grouping input letters because of its ambiguity. As for JSM, sequence segmentation utilises one-to-one alignment. An example of the segmentation is shown in Figure 4.

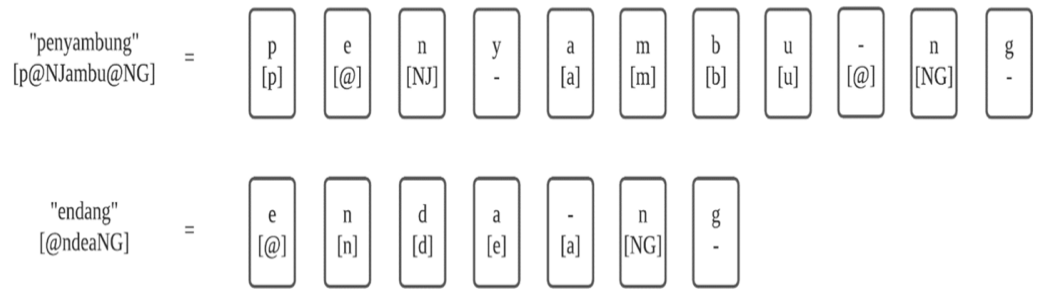

Figure 4. Segmentation sequence according to JSM

Hence, the joint probability of obtaining the sequence is determined by the total of all matching graphone sequences (Equation 15):

$$p(g, \varphi) = \sum_{q \epsilon S(g, \varphi)} p(\boldsymbol{q}) \tag{15}$$

where $q \epsilon Q^*$ is a sequence of graphones and $S(g, \varphi)$ contains all cosegmentations of $\boldsymbol{g}$ and $\boldsymbol{\varphi}$ (Equation 16).

$$S(g, \varphi) \coloneqq \left\{ q \epsilon Q^* \middle| \begin{matrix} g_{q_1} \frown \cdots \frown g_{q_K} = g \\ \varphi_{q_1} \frown \cdots \frown \varphi_{q_K} = \varphi \end{matrix} \right\} \tag{16}$$

Equation 16 ⌣ denotes the sequence concatenation, and K = |*q*| is the sequence length by ***q***. The joint probability can be modelled using a standard M-gram approximation as Equation 17:

$$p(q_1^K) \cong \prod_{j=1}^{K+1} p\left(q_j \middle| q_{j-1}, \ldots, q_{j-M+1}\right) \quad (17)$$

where the positions $j < 1$ and $j > K$ have a boundary that allows modelling characteristic phenomena at the start and end of the word.

A few numbers of studies have been using these methods for grapheme-to-phoneme conversions, such as the study of Saychum et al. (2016), Wang and Sim (2013), and Masmoudi et al. (2016).

Table 3 shows the summary of all reviewed G2P techniques. The summary includes the name of the techniques, the overall working process and the studies that have adopted the techniques in their experiments or journals.

Table 3

Summary of G2P techniques

G2P technique	How does it work	Studies that used the technique
CRF	Using a network of non-directional nodes and vertices	Illina et al. (2011), Zweig and Nguyen (2009), Yamazaki et al. (2014), Al-Shareef and Hain (2012), Masmoudi et al. (2016)
JMM	Segments of the word letter sequence in the many-to-one alignment	Cherifi and Guerti (2021), Masmoudi et al. (2016), Wang and Sim (2013)
JSM	Segments the word letter sequences using one-to-one alignment	Saychum et al. (2016), Wang and Sim (2013), Masmoudi et al. (2016)

G2P for Building Pronunciation Variants

As discussed in the earlier part of this paper, pronunciation variants are important, as shown by the English language, commonly known as British and American English. Including pronunciation variants in the system helps to show that another form of intonation and spelling exists for the words. Most rich-resource languages have included pronunciation variants (dialect, slang) in the system; including the variants is less challenging for this category, as their resources are ample.

However, such is not the case for under-resourced language. In the last few years, some studies focused on under-resourced language, which deserved the recognition of

researchers and developers. However, the efforts of building a G2P system for under-resourced languages vary, such as bootstrapping (Juan & Flora, 2015), web-mining, and segmentation (Saychum et al., 2016; Tsuboi et al., 2008).

Only one study investigated G2P for building pronunciation variants where the target language comes from an under-resourced language. Such analysis has been described by (Lukeš et al., 2019), where the study used the Czech language from two sources to generate pronunciation variants. The number of studies that are alike is very scarce. There is still little work on G2P modelling to produce pronunciation variants. Thus, there is a need to study the computational approach for this research gap. By following the example of the study mentioned, it is possible to use an under-resourced language, such as the Iban language, as the target language to generate pronunciation variants. Thus, it needs to include the variants in the pronunciation dictionary systematically.

Under-Resource Language and G2P Challenges

Joshi et al. (2020) mention that languages worldwide are grouped according to classes. There are six (6) class 'races', as shown in Table 4.

From Table 4, under-resourced languages fall in positions 0, 1, and 2, and a few managed to get into position 3. The number of speakers from these positions ranges from 1.8 billion and below (Joshi et al., 2020); meanwhile, the number of speakers for position 4 and 5 are 2.0 billion and above. Languages in these positions are known as 'rich-resourced languages', such as Russian, Korean, English, Spanish, German, Japanese, and French.

Under-resourced languages are almost endangered because they are fewer studied and digitalised records regarding the language (Singh, 2008). They are also losing to extinction as the native under-resourced speakers resorted to using high-resource languages such as English and French. Furthermore, fewer and fewer people can talk using under-resourced language, other than a very little guide to that language in the community themselves (Brenzinger et al., 2003).

Table 4
Positions and class 'race' of languages

Position	Class 'race'
0	The Left-Behinds
1	The Scraping-Bys
2	The Hopefuls
3	The Rising Stars
4	The Underdogs
5	The Winners

Some prime examples of under-resourced languages have been stated in the introduction of this paper. In addition, studies focus on these under-resourced languages and how to preserve them. Still, there are limitations in pursuing the conservation of the languages-insufficient amount of data, fewer language experts on the language, and limitations of the software used to emulate the experiments.

These limitations are also applied in G2P conversion and modelling techniques. An immense amount of data is needed for the experiments (training and testing). Since under-resourced languages have limited data, some researchers such as Juan and Flora (2015), Juan & Besacier (2013) and Juan et al. (2015) use an approach where the language is closely related to the under-resourced language that is being used for training. The process, known as bootstrapping, has generated the baseline dictionary for the under-resourced language. However, the process consumed much time as it started the baseline from scratch and depended on the equipment used.

As Besacier et al. (2014) stated, the challenge of bridging the gap between language and technology experts still holds until recent years. Language experts regarding under-resourced languages are rare and mainly originated outside the said language's country. Only a few pursue such language and give effort to researching and including them in language technology. Efforts are made because they realise the danger of language extinction and how it affects the world languages, which many people, especially native speakers, do not. It is also quite rare to find a language expert with the knowledge of developing ASR that is native to the under-resourced language itself.

DISCUSSION

As discussed earlier regarding G2P conversion techniques, three G2P modelling techniques are available: CRF, JMM and JSM. All three modelling techniques have been used in G2P conversion for rich and under-resourced languages.

CRF modelling technique has non-directional nodes and vertices, making the training process in the conversion have more match-up probability and a wide selection of letter-phone pairing. Since CRF is an upgrade from HMM and MEMM, the modelling technique has solved labelling and observation biases (McCallum, 2012). CRF is said to be independent and flexible in creating segmentations if the selected feature is correct. However, this can also invite unwanted errors in the generated output. It is mentioned to have high computational complexity during the algorithm's training stage, making it harder for model re-training when new data is included. It is also difficult for CRF training to detect and learn unknown words not included in the training data.

JSM and JMM look almost the same, but there is a slight difference during the segmentation process. JSM segmentation alignment focuses on one-to-one alignment. Meanwhile, JMM utilises many-to-any alignments. However, the main concern for both modelling techniques is the sparseness problem (Bisani & Ney, 2002). Each word's various sizes or lengths can yield different results; features such as evidence trimming and maximum approximation (Viterbi training) are important in segmentation. The features must be balanced and used when appropriate to yield a better result.

Another term that has been mentioned before is the term bootstrapping, or bootstrap. It is "to change the state using existing resources" (Zoubir & Iskander, 2007). As a data-driven approach, this method can substitute tedious and often impossible analytical deviations with computational analysis and calculations. For bootstrapping to succeed, the most suitable resampling schemes must be identified. Initial decisions must be based on examining the data and the problem. Next, determine whether the data in hand is independently and identically distributed (IID in short) or non-IID. IID data is the collection of data that are unlikely to happen in an actual situation, but for study simulations, it is sufficient. If the data is identified as IID, standard bootstrap resampling techniques like independent data bootstrap can be used. If it is non-IID, consider using a parametric approach where a specific structure is assumed, which helps reduce the problem difficulty of dependent data bootstrap to standard resampling of the assumed IID model error estimates. The flow of the strategy process can be seen in Figure 5.

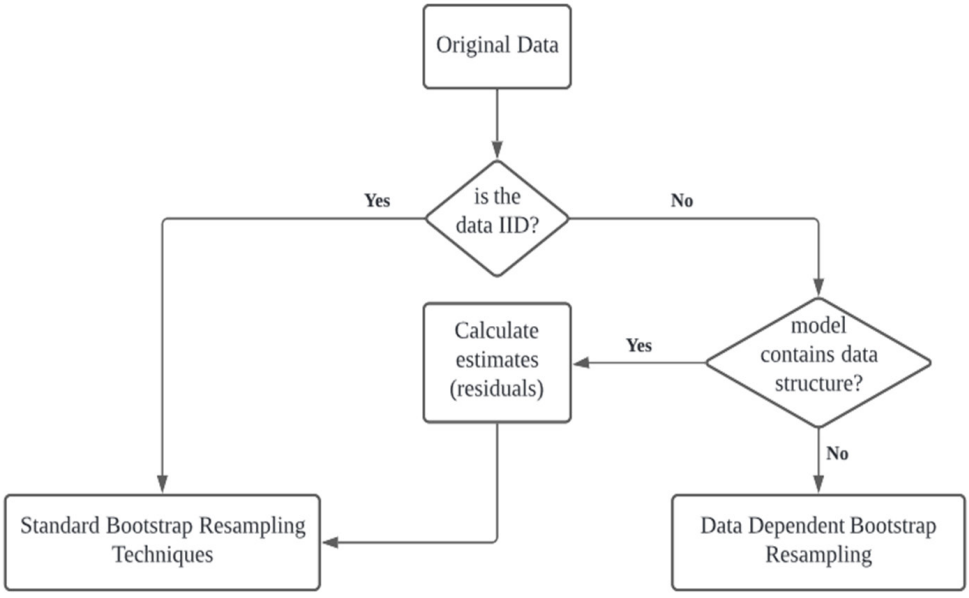

Figure 5. Practical strategy for bootstrapping data (Zoubir & Iskander, 2007)

The bootstrapping method can be applied to estimate statistical characteristics such as bias, variance, distribution functions and, thus, confidence intervals. The method itself is a computational tool for statistical inference. Furthermore, the bootstrap method can estimate hypothesis tests and model selection (Zoubir & Iskander, 2007). Zoubir and Iskander (2007) state that the bootstrap method can be used in experiments with very little data on hand to avoid using asymptotic results. In terms of G2P, the bootstrap method is being used in a data-driven approach, where the estimation and substitution to the unknown while

referring to the available data. One example has been portrayed in the study of Juan and Besacier (2013), where it has been done by filling and substituting unknown data into the Iban pronunciation dictionary using a closely-related language, Malay.

Out-of-Vocabulary, or OOV in short, is a term for words of the selected language in the dictionary that is not a part of the dictionary list of words. It is one of the causes of error in recognising spontaneously spoken utterances (Young, 1994).

The modelling techniques mentioned are based on a data-driven approach, which means the learning or training process uses the input seed data as a reference. The data-driven approach can be semi-supervised or unsupervised most of the time since no default rules are needed in the learning process. Other than that, a data-driven approach may be the best option for generating pronunciation variations. Especially in under-resourced language, when there is scarce information regarding linguistic information (Amdal et al., 2000). In the case of the Iban language, the documentation of rules and exceptions of the language structure is incomplete. Hence, the best approach is to use closely related language data as the seed data for training and testing. Based on the study made by Amdal et al. (2000), the pattern of the data-driven approach in any modelling technique usually follows these steps in G2P conversion:

- Automatically generate alternative transcription.
- Align reference and alternative transcription
- Derive the initial rule from the alignment

However, a data-driven approach can also be rule-based, like a knowledge-based approach that uses data to learn and modify pre-determined pronunciation rules.

CONCLUSION

Pronunciation dictionary plays an important role in ASR systems, especially for an under-resourced language such as the Iban language. It helps to show the variation of word spellings and language pronunciation. Culturally, the Iban language has its pronunciation variants, varying across Sarawak's region. However, the variants are not explicitly depicted in the Iban pronunciation dictionary developed for ASR tasks.

Currently, there is no efficient way of generating Iban pronunciation variants; thus, this paper reviewed selected G2P methods for generating pronunciation variants. Furthermore, this paper reviewed methods used for G2P conversions, such as CRF, JMM and JSM. These methods produced pronunciation variants by converting grapheme sequences to phoneme sequences from any target language.

We described the importance of pronunciation variants, particularly for ASR systems. The variants in the pronunciation dictionary can increase the ASR system's probability of analysing word variations and the base word of the targeted language, thus enriching the system's lexicon with various possibilities. Thus, there is a need to develop systematic

approaches to include significant variants in the dictionary to cater to speakers' speech variabilities when using ASR applications. Articles highlighted in this paper can bring insights to researchers on recent works. These works are related to developing pronunciation variants for under-resourced languages and state-of-the-art techniques for producing G2P models that are reliable in predicting pronunciations for Out-Of-Vocabulary (OOV) words.

We have also briefly discussed the bootstrap method and its correlation with the data-driven approach. The discussion has included the bootstrap application, how it works in a data-driven environment, and the practical bootstrapping strategy in G2P techniques, as shown in Figure 5. Moreover, this paper can be used as a guide or a baseline study on G2P modelling pronunciation variants for under-resourced languages.

ACKNOWLEDGEMENTS

This work was supported by the Malaysia Comprehensive University Network (Grant number: GL/F08/MCUN/ 11/2020). We are grateful for the support enabling us to conduct this research.

REFERENCES

Al-Shareef, S., & Hain, T. (2012). Crf-based diacritisation of colloquial Arabic for automatic speech recognition. In *Thirteenth Annual Conference of the International Speech Communication Association* (pp. 1824-1827). ISCA Publishing.

Amdal, I., Korkmazskiy, F., & Surendran, A. C. (2000, October 16-20). Joint pronunciation modelling of non-native speakers using data-driven methods. In *INTERSPEECH* (pp. 622-625). Beijing, China.

Besacier, L., Barnard, E., Karpov, A., & Schultz, T. (2014). Automatic speech recognition for under-resourced languages: A survey. *Speech Communication*, *56*(1), 85-100. https://doi.org/10.1016/j.specom.2013.07.008

Bisani, M., & Ney, H. (2002, September 16-20). Investigations on joint-multigram models for grapheme-to-phoneme conversion. In *INTERSPEECH* (pp. 1-4). Colorado, USA

Bisani, M., & Ney, H. (2008). Joint-sequence models for grapheme-to-phoneme conversion. *Speech Communication*, *50*(5), 434-451. https://doi.org/10.1016/j.specom.2008.01.002

Brenzinger, M., Yamamoto, A., Aikawa, N., Koundiouba, D., Minasyan, A., Dwyer, A., Grinevald, C., Krauss, M., Miyaoka, O., Sakiyama, O., Smeets, R., & Zepeda, O. (2003, March 10-12). Language vitality and endangerment. In *International Expert Meeting on the UNESCO Programme Safeguarding of Endangered Languages*. Fontenoy, Paris.

Chen, S., Beeferman, D., & Rosenfeld, R. (1998, February 8-11). Evaluation metrics for language models. In *Proceedings of the DARPA Broadcast News Transcription and Understanding Workshop* (pp. 275-280). Lansdowne, Virginia. http://repository.cmu.edu/cgi/viewcontent.cgi?article=2330&context=compsci

Cherifi, E. H., & Guerti, M. (2021). Arabic grapheme-to-phoneme conversion based on joint multi-gram model. *International Journal of Speech Technology*, *24*(1), 173-182. https://doi.org/10.1007/s10772-020-09779-8

Chowdhury, S. A., Alam, F., Khan, N., & Noori, S. R. H. (2018). Bangla grapheme to phoneme conversion using conditional random fields. In *2017 20th International Conference of Computer and Information Technology (ICCIT)* (pp. 1-6). IEEE Publishing. https://doi.org/10.1109/ICCITECHN.2017.8281780

Deligne, S., Yvon, F., & Bimbot, F. (1995, September 18-21). Variable-length sequence matching for phonetic transcription using joint multigrams. In *Fourth European Conference on Speech Communication and Technology* (pp. 2243-2246). Madrid, Spain.

Guazzi, M. D., Cipolla, C., Sganzerla, P., Agostoni, P. G., Fabbiocchi, F., & Pepi, M. (1983). Language vitality and endangerment. *European Heart Journal*, *4*(Suppl. A), 181-187. https://doi.org/10.1093/eurheartj/4.suppl_a.181

Illina, I., Fohr, D., & Jouvet, D. (2011, August 28-31). Grapheme-to-phoneme conversion using Conditional Random Fields. In *Twelfth Annual Conference of the International Speech Communication Association* (pp. 2313-2316). Florence, Italy.

Joshi, P., Santy, S., Budhiraja, A., Bali, K., & Choudhury, M. (2020). *The state and fate of linguistic diversity and inclusion in the NLP world*. arXiv Preprint. https://doi.org/10.18653/v1/2020.acl-main.560

Juan, S., & Flora, S. (2015). *Exploiting resources from closely-related languages for automatic speech recognition in low-resource languages from Malaysia* (Doctoral dissertation). Université Grenoble Alpes, France. https://www.theses.fr/2015GREAM061

Juan, S. S., & Besacier, L. (2013, October 14-18). Fast bootstrapping of grapheme to phoneme system for under-resourced languages-application to the iban language. In *Proceedings of the 4th Workshop on South and Southeast Asian Natural Language Processing* (pp. 1-8). Nagoya, Japan.

Juan, S. S., Besacier, L., Lecouteux, B., & Dyab, M. (2015, September 6-10). Using resources from a closely-related language to develop ASR for a very under-resourced language: A case study for iban. In *Proceedings of the Annual Conference of the International Speech Communication Association* (pp. 1270-1274). Dresden, Germany.

Jurafsky, D., & Martin, J. (2000). *Speech & Language Processing*. Pearson Education India.

Karanasou, P. (2013). *Phonemic variability and confusability in pronunciation modeling for automatic speech recognition* (Doctoral dissertation). Université Paris Sud-Paris, France. http://hal.archives-ouvertes.fr/tel-00843589/

Lafferty, J., McCallum, A., & C.N. Pereira, F. (2001). Conditional rndom fileds: Probabbilistic models for segmenting and labeling sequence data. In *Proceedings of the 18th International Conference on Machine Learning 2001 (ICML 2001)* (pp. 282-289). ACM Publishing. https://doi.org/10.29122/mipi.v11i1.2792

Laurent, A., Meignier, S., & Deléglise, P. (2014). Improving recognition of proper nouns in ASR through generating and filtering phonetic transcriptions. *Computer Speech & Language, 28*(4), 979-996. https://doi.org/10.1016/j.csl.2014.02.006

Lukeš, D., Kopřivová, M., Komrsková, Z., & Poukarová, P. (2018, May 7-12). Pronunciation variants and ASR of colloquial speech: A case study on Czech. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)* (pp. 2704-2709). Miyazaki, Japan.

Masmoudi, A., Ellouze, M., Bougares, F., Esètve, Y., & Belguith, L. (2016). Conditional random fields for the tunisian dialect grapheme-to-phoneme conversion. In *Proceedings of the Annual Conference of the International Speech Communication Association* (pp. 1457-1461). ISCA Publishing. https://doi.org/10.21437/Interspeech.2016-1320

McCallum, A. (2012). *Efficiently Inducing Features of Conditional Random Fields*. arXiv Preprint. http://arxiv.org/abs/1212.2504

Morris, J. J. (2010). *A study on the use of conditional random fields for automatic speech recognition* (Doctoral dissertation). The Ohio State University, USA. https://etd.ohiolink.edu/apexprod/rws_olink/r/1501/10?clear=10&p10_accession_num=osu1274212139

Omar, A. (1981). *The Iban language of Sarawak; A grammatical description*. Kuala Lumpur: Dewan Bahasa dan Pustaka.

Ramli, I., Jamil, N., Seman, N., & Ardi, N. (2015). An improved syllabification for a better Malay language text-to-speech synthesis (TTS). *Procedia Computer Science*, *76*, 417-424. https://doi.org/10.1016/j.procs.2015.12.280

Rugchatjaroen, A., Saychum, S., Kongyoung, S., Chootrakool, P., Kasuriya, S., & Wutiwiwatchai, C. (2019). Efficient two-stage processing for joint sequence model-based Thai grapheme-to-phoneme conversion. *Speech Communication*, *106*, 105-111. https://doi.org/10.1016/j.specom.2018.12.003

Saychum, S., Kongyoung, S., Rugchatjaroen, A., Chootrakool, P., Kasuriya, S., & Wutiwiwatchai, C. (2016, September 8-12). Efficient Thai grapheme-to-phoneme conversion using CRF-based joint sequence modeling. In *Proceedings of the Annual Conference of the International Speech Communication Association* (pp. 1462-1466). ISCA Publishing. https://doi.org/10.21437/Interspeech.2016-621

Shin, C. (2021). Iban as a koine language in Sarawak. *Wacana*, *22*(1), 102-124. https://doi.org/10.17510/wacana.v22i1.985

Singh, A. K. (2008). Natural language processing for less privileged languages: Where do we come from? Where are we going? In *Proceedings of the IJCNLP-08 Workshop on NLP for Less Privileged Languages* (pp. 7-12). Asian Federation of Natural Language Processing. http://www.aclweb.org/anthology/I08-3004

Stadtschnitzer, M., & Schmidt, C. (2018, May 7-12). Data-driven pronunciation modeling of swiss german dialectal speech for automatic speech recognition. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)* (pp. 3152-3156). Miyazaki, Japan.

Sutlive, V. H. (1994). *A handy Reference Dictionary of Iban and English*. Tun Jugah Foundation.

Tjalve, M., & Huckvale, M. (2005, September 4-8). Pronunciation variation modelling using accent features. In *9th European Conference on Speech Communication and Technology* (pp. 1341-1344). Lisbon, Portugal.

Tsuboi, Y., Kashima, H., Mori, S., Oda, H., & Matsumoto, Y. (2008, August 18-22). Training conditional random fields using incomplete annotations. In *Coling 2008 - 22nd International Conference on Computational Linguistics, Proceedings of the Conference* (pp. 897-904). Manchester, UK. https://doi.org/10.3115/1599081.1599194

Tan, T. P., Xiao, X., Tang, E. K., Chng, E. S., & Li, H. (2009). MASS: A Malay language LVCSR corpus resource. In *2009 Oriental COCOSDA International Conference on Speech Database and Assessments* (pp. 25-30). IEEE Publishing. https://doi.org/10.1109/ICSDA.2009.5278382.

Wang, X., & Sim, K. C. (2013). Integrating conditional random fields and joint multi-gram model with syllabic features for grapheme-to-phone conversion. In *INTERSPEECH* (pp. 2321-2325). ISCA Publishing.

Yamazaki, M., Morita, H., Komiya, K., & Kotani, Y. (2014). Extracting the translation of anime titles from web corpora using CRF. In *Knowledge-Based Software Engineering: 11th Joint Conference, JCKBSE 2014* (pp. 311-320). Springer International Publishing. https://doi.org/10.1007/978-3-319-11854-3_26

Yolchuyeva, S., Németh, G., & Gyires-Tóth, B. (2019). Grapheme-to-phoneme conversion with convolutional neural networks. *Applied Sciences*, *9*(6), 1-17. https://doi.org/10.3390/app9061143

Young, S. R. (1994, April). Detecting misrecognitions and out-of-vocabulary words. In *Proceedings of ICASSP'94. IEEE International Conference on Acoustics, Speech and Signal Processing* (Vol. 2, pp. II-21). IEEE Publishing. https://doi.org/10.1109/ICASSP.1994.389728

Yu, M., Nguyen, H. D., Sokolov, A., Lepird, J., Sathyendra, K. M., Choudhary, S., Mouchtaris, A., & Kunzmann, S. (2020). Multilingual grapheme-to-phoneme conversion with byte representation. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 8234-8238). IEEE Publishing. https://doi.org/10.1109/ICASSP40776.2020.9054696

Zoubir, A. M., & Iskander, D. R. (2007). Bootstrap methods and applications : A tutorial for the signal processing practitioner. *IEEE Signal Processing Magazine*, *24*(4), 10-19. https://doi.org/10.1109/MSP.2007.4286560

Zweig, G., & Nguyen, P. (2009). Maximum mutual information multi-phone units in direct modeling. In *Tenth Annual Conference of the International Speech Communication Association* (pp. 1919-1922). ISCA Publishing.