SEMI-SUPERVISED G2P BOOTSTRAPPING AND ITS APPLICATION TO ASR FOR A VERY UNDER-RESOURCED LANGUAGE: IBAN

Sarah Samson Juan, Laurent Besacier, Solange Rossato

{sarah.samson-juan, laurent.besacier, solange.rossato}@imag.fr
Grenoble Informatics Laboratory (LIG)
University Grenoble-Alpes
Grenoble, France

ABSTRACT

This paper describes our experiments and results on using a local dominant language in Malaysia (Malay), to bootstrap automatic speech recognition (ASR) for a very under-resourced language: Iban (also spoken in Malaysia on the Borneo Island part). Resources in Iban for building a speech recognition were nonexistent. For this, we tried to take advantage of a language from the same family with several similarities. First, to deal with the pronunciation dictionary, we proposed a bootstrapping strategy to develop an Iban pronunciation lexicon from a Malay one. A hybrid version, mix of Malay and Iban pronunciations, was also built and evaluated. Following this, we experimented with three Iban ASRs; each depended on either one of the three different pronunciation dictionaries: Malay, Iban or hybrid. Our best results (WER) for Iban ASR (with different lexicon) were as follows: 20.82% (Malay G2P), 21.90% (Iban G2P) and 20.60% (Hybrid G2P). Apart from that, we applied system combination using all of the systems and obtained an improved accuracy of 19.22%.

Index Terms— under-resourced language, speech recognition, Iban language, Malay language, bootstrapping, Kaldi, grapheme-to-phoneme

1. INTRODUCTION

Phonetic lexicons are crucial for speech applications and the process for creating one for a new language can take a significant amount of time and effort. This is due to the fact that such lexicons are not readily available for these languages. Manual Grapheme-to-phoneme (G2P) conversion is obviously not an option to get several thousands of pronunciations from a word list. However, there are data driven techniques to help train G2P systems, for example [1], [2] and [3], where primarily, a base phonetic dictionary is required for training. The pronunciation model then, can be used to decode new words (OOV words) to phoneme sequences, limited to the predefined phoneme set.

Bootstrapping G2P has been implemented to assist in creating pronunciation lexicons for languages such as Afrikaans and Nepali ([4], [5]). This semi-supervised method predicts additional entries of a dictionary through a pronunciation model and the outputs are then verified by a native speaker or linguist. Typically, a seed lexicon in the target language must be prepared initially for this purpose. More often than not, knowledge on new languages are poor. Hence, it is a constant challenge in generating this data for under-resourced languages [6].

Our paper introduces a feasible approach that is suitable for languages from the same language family. We propose to use Malay pronunciations to produce a base phonemic transcript for Iban and then, post-edit the outputs to match Iban pronunciations. The idea rises due to the fact that Malay and Iban are closely related (same writing system, phonetically similar) and they belong to the same language branch. In this paper, we briefly describe our investigation over Malay and Iban pronunciation distance using the access data that we currently have. Heeringa and de Wet in [7] conducted a similar study to measure the distances between Afrikaans and Dutch, Afrikaans and Frisian; as well as Afrikaans and German based on phoneme transcripts. By using this method, the authors concluded that Dutch has pronunciations closer to Afrikaans than to the other two languages.

After generating and post-editing the transcripts, we then train an Iban G2P system to phonetize other entries. Both Malay and Iban G2Ps are evaluated, after which, results suggesting a strategy for converting the whole Iban vocabulary. The remainder of the paper explains further in detail about our investigation and experiments to evaluate Iban ASR. Section 2 describes briefly about Iban and Section 3 reports collected resources for Iban ASR and the strategy that we employ for building G2P. Section 4 presents acoustic model types for Iban, decoding strategies and results of three Iban ASRs. Section 5 provides information about the outputs generated by the three recognizers and last but not least, Section 6 concludes the paper and gives perspectives.

Table 1. *Malay/Iban examples with their pronunciations*

No.	Word [meaning(s) I : M]	Iban	Malay
1	ke [I=M : to]	/kə/	/kə/
2	nya [that : him/her]	/ɲaʔ/	/ɲə/ or / ɲa/
3	kayu [I=M: wood]	/kajuʔ/	/kaju/
4	bilik [I=M: room]	/biliəʔ/	/bileʔ/
5	dua [I=M: two]	/duwa/	/duwə/ or /duwa/
6	kepala [head,leader : head]	/kepalaʔ/	/kepala/
7	puluh [I:M : -ty (quantity)]	/puluəh/	/puloh/
8	raban [group : rambling speech, talk rapidly]	/raban/	/raban/
9	lalu [then, pass-by : then, before (time), unwell]	/lalu/	/lalu/
10	orang [I=M : person, people]	/uraŋ/	/oraŋ/

I : Iban, M : Malay, I = M : same meaning in Malay and Iban

2. THE IBAN LANGUAGE

Iban[1] is a member of the Austronesian language family, under the Ibanic group. The language belongs to the Malayo-Sumbawan[2] branch as Malay, where, the latter is under to the Malayan group ([8], [9], [10]).With over 600,000 Iban speakers, the language is mostly spoken in Sarawak, Kalimantan and Brunei. In the course of modernization, there are also Iban speakers found in the Peninsula Malaysia particularly in Johor and Kuala Lumpur. Alongside learning Malay (the official language of Malaysia), Iban has been taught in schools at the primary and secondary levels and it is offered as a nonobligatory subject since the early 90s. In several universities in Malaysia, beginner level courses are offered to undergraduate students to attract non-native speakers learn Iban.

The Iban system is influenced by the Malay system in terms of phonology, morphology and syntax [11]. According to a guidebook on Iban [11], there are common words (same surface forms) between the two languages and also, there are Malay words integrated (borrowed) to the Iban language. Example of Malay / Iban same surface forms and their pronunciations are listed in Table 1. In 1981, Omar [12] published a complete description of the language. In her work, she included phoneme classifications and morphological details of Iban. According to the author, there are 19 consonants (/p/, /b/, /m/, /w/, /t/, /d/, /n/,/ʧ/, /ʤ/, /s/, /l/, /r/, /ɲ/, /j/, /k/, /g/, /ŋ/, /h/, /ʔ/), 6 vowels (/a/, /e/, / ə/, /i/, /o/, /u/) and 11 vowel clusters (/ui/, /ia/, /ea/, /ua/, /oa/, /iu/, /iə/, /uə/, /oə/, /ai/, /au/). This list of consonants did not include consonants from Malay such as /f/, /v/, /θ/, /z/, /x/, /ɣ/, /ð/, /ʃ/ which frequently appear in loan words. As shown in Table 1, Iban and Malay orthographies are Latin based where both use 26 English alphabets. Furthermore, Iban and Malay are non tonal languages. The obvious differences between Malay and Iban are the presence of vowel clusters between two consonants and /ʔ/ sound for some words ending with a vowel.

3. IBAN RESOURCES

3.1. Text data for language modelling

Iban electronic texts were found and used for this study. News data was collected from a news website[3] that produces Iban articles daily. We crawled articles from 2009 to 2012 and we succeeded in gathering a total of 7K news articles concerning general, sports and entertainment. After the extraction, the text was cleaned and normalized by : (1) removing HTML tags, (2) converting dates and numbers to words (e.g: 1982 to *sembilan belas lapan puluh dua*), (3) converting abbreviations to full terms (e.g: Dr. to *Doktor*, Prof. to *Profesor*, Kpt. to *Kapten*), (4) splitting paragraphs to sentences, (5) changing uppercase characters to lowercase and (6) removing punctuation marks (except hyphen / '-'). Finally, we have approximately 2.08M words for our experiments.

Using this text data, we built a trigram Iban language model with modified Kneser-Ney discounting. SRILM [13] toolkit was used to obtain the model and later, measured the model's perplexity on Iban speech transcripts (see next section for the speech corpus). The evaluation gave us a perplexity of 158 and 2.3% OOV rate.

3.2. Speech corpus and transcript

We have eight hours of news data with 16 kHz sampling rate. The data was provided by the Radio Televisyen Malaysia (RTM), a local radio and television station, through one of its channel, Waifm. The channel airs five to ten minutes of news in Iban daily. Table 2 and Table 3 list the database and our experiment setting.

[1]Iban language code : [iba] (ISO 639-3)

[2]For details : http://wals.info/langoid/genus/malayosumbawan#4/5.10/118.08

[3]www.theborneopost.com/news/utusan-borneo/berita-iban/

Table 2. Iban speech corpus statistics

Gender	Speakers	Sentences	Tokens	Length (mins)
Female	14	1,382	36,194	222
Male	9	1,750	44,408	257

Table 3. Train / test splits for experiments

Set	Speakers	Gender (M:F)	Sentences	(mins)
Testing	6	2:4	473	71
Training	17	7:10	2659	408

The speech data was transcribed by eight Iban native speakers including seven female. Prior to completing their tasks, the transcribers were given a training session on Transcriber ([14], [15]), an open source tool for segmenting, labeling, and transcribing speech. The tool assists them in annotating noise (music, page turns, etc) and utterances as well as segmenting signals to separate multiple sentences. In total, there are 3,132 sentences uttered by 25 speakers and 473 sentences were chosen for evaluation that last a little over an hour.

3.3. Pronunciation dictionary

3.3.1. Obtaining the Malay G2P

First, we obtained a Malay pronunciation lexicon from Tan et al. [16]. The authors developed a 76K pronunciation lexicon for their Malay speech recognition that eventually gave them an ASR baseline result of 14.6% WER. Using their lexicon, we trained a Malay grapheme to phoneme (G2P) as a base G2P system for this study. Training was done on Phonetisaurus, an open source tool based on Weighted Finite States Transducers ([17], [1]). The training size was 68K and the phonetizer was evaluated using 8K Malay data. The results were 6.20% phoneme error rate (PER) and 24.98% WER (refer to [18] for more details on the experiment setup).

3.3.2. Obtaining the Iban G2P

We found 37K unique words in the Iban text data. The list has Malay (23%) and English (19%) words, a verdict we made after conducting a comparison study using Malay (from [16]) and English (CMU version for Sphinx) vocabularies. Following that, we were intrigued to know about the pronunciation distances between Malay and Iban, especially for the same surface forms (hereafter, we address as Malay-Iban). To implement this, we applied Levenshtein [19] method to calculate the distances, following a similar study conducted by [7] where they measured pronunciation distances between Afrikaans and Dutch, Afrikaans and Frisian; as well as Afrikaans and German. By using this method, the authors were able to conclude that Dutch has closer pronunciations to Afrikaans compared to the other two languages based on the phonemic transcripts.

We tested on 100 most frequent Malay-Iban words found in the Iban text data. Phoneme sequences for Malay were generated using Malay phonetizer and then the outputs were post-edited to match Iban pronunciations. The latter part was done by an Iban native speaker[4]. The correction involved inserting glottal stops and substituting/inserting vowels. We fixed to use Malay phoneme set only and for that reason no new phonemes created at this point. Then, we evaluated the post-edited version with the Malay one and found that we obtained 17% PER and 47% WER, which indicates that only 53 pronunciation pairs were equivalent (no change). Based on these results, we were motivated to continue to apply this semi-supervised approach to the rest of data and analyze the consequences.

Table 4. The Malay and Iban phonetizers performances for an Iban phonetization task

Phonetizer	Corpus	PER (%)	WER (%)	Post-edit (mins)
Malay G2P	500_{IM}	6.52	27.2	30
	500_{I}	15.8	56.0	42
Iban G2P	500_{IM}	13.6	44.2	45
	500_{I}	8.2	31.8	32

IM for Malay-Iban words and I for pure Iban words. [18]

Hence, we phonetized 1K words that consist of 500 Malay-Iban and 500 pure Iban (not shared with neither Malay nor English orthography) words. Sequences were post-edited and we trained our first Iban G2P using this data. Later, the two phonetizers were evaluated to measure performance of phonetizing Iban words. To do this, we evaluated another 1K words (same protocol as before) and our results in [18] showed that Malay G2P can phonetize Malay-Iban better than pure Iban, whereas Iban G2P works better for pure Iban (see Table 4 for a summary of results).

3.3.3. Obtaining pronunciations for the whole lexicon

After analyzing the Malay and Iban G2P performances separately, we decided to generate pronunciations for Iban using both phonetizers. The strategy was as follows: the Malay G2P phonetizes all Malay-Iban while the Iban G2P phonetizes all pure Iban words. Besides that, we also apply Malay G2P to English words that are found in the Iban lexicon. This is because the phonetizer is able to phonetize English as demonstrated in [16]'s work for Malay recognizer. Using this proposed training strategy, we have 37K pronunciations including 1K of Iban G2P data. The

[4]the first author of this paper

pronunciation lexicon is estimated to have 8.1% PER and 29.4% WER on 2K random outputs.

3.3.4. Analyzing pronunciations

In addition to having a mix of Malay and Iban pronunciations in the dictionary (later address as Hybrid G2P), we generated two other pronunciation lexicons. One has Malay pronunciations, which we obtained after employing the Malay G2P to the whole Iban lexicon and the second list has Iban pronunciations generated by the Iban phonetizer (1K).

A comparison study was carried out to analyze the pronunciation dictionaries and our findings are presented in Table 5. Here, we denote the phonetizers using the following labels for simplicity: Malay G2P as $S1$, Iban G2P as $S2$ and Hybrid G2P as $S3$. Let $C_{\mathbf{AB}}$ has elements that are **not** common to $\mathbf{A}$ and $\mathbf{B}$ From Table 5, 67% of pronunciations obtained with Malay G2P are different than those obtained with Iban G2P. Meanwhile, the hybrid version (S3) is closer to Iban G2P (S2) with 29% error.

Table 5. Comparison results between two pronunciation dictionaries (total words 36K)

$C_{\mathbf{AB}}$	No. of diff. pronunciations	%
C_{S1S2}	24,587	67.6
C_{S1S3}	14,162	39.0
C_{S2S3}	10,593	29.1

We investigated further to determine which language group lead to the words with different pronunciations (elements of $C_{\mathbf{AB}}$). As mentioned in Section 3.3.2, there are Malay and English words in the Iban lexicon. Therefore, we categorized according to three groups, English, Malay and the rest as pure Iban. We present the results as in Table 6. When we compared S1 (Malay G2P) to S2 (Iban G2P), majority of the differences belong to pure Iban and the same can be said after we compared Malay G2P to Hybrid G2P. As English words were phonetized by S3 (Iban G2P) system, their pronunciations are different with S1 (Malay G2P) and S3 (Hybrid G2P) while no differences are seen between Malay G2P and Hybrid G2P for these. The reason is that English pronunciations were already available in our Malay lexicon.

Table 6. Statistics of words in Table 5 according to three language groups

Language	C_{S1S2}	C_{S1S3}	C_{S2S3}
English	5,605	0	5,605
Malay	5,031	202	4,912
pure Iban	13,951	13,960	76

4. BASELINE SPEECH RECOGNIZERS

We experimented Kaldi ASR system [20] for Iban, an open source toolkit based on Finite States Transducers. Acoustic models were trained using three lexicons and the training transcript. Each system is called Malay G2P, Iban G2P or Hybrid G2P, depending on which lexicon is applied. We explored several techniques offered by Kaldi for training the systems. We extracted 13 MFCCs and Gaussian mixture models were employed for monophone and triphone trainings. For the triphone, we applied 2,998 context-dependent states and 40,000 Gaussians. We also implemented delta delta coefficients on the MFCCs, linear discriminant analysis (LDA) transformation and maximum likelihood transform (MLLT) [21], and, speaker adaptation based on feature-space maximum likelihood linear regression (fMLLR) [22]. Then, decoding was launched by applying language model scales of 5.0 to 20.0 thus, resulting 16 WERs per decoding. We report results for the best of them in the next section.

4.1. ASR Results

The baseline results are summarized in Table 7. Results obtained using monophone models provide us an average of 42% WER. Gradually, the accuracies increased as triphone models were used and different features employed. The final results brought an average of 21% WER, a half of the average result using monophone models. Surprisingly, the difference between the three recognizers' performances are not much. Our best results are merely 1% difference between each and among the three recognizers, the system with a mix of Malay and Iban pronunciations is the best one (20.6% WER).

Table 7. Iban recognizers performances (WER%) based on different approaches applied

Training approach	Dictionary		
	Malay G2P	Iban G2P	Hybrid G2P
Monophone	42.17	41.79	41.97
Triphone	36	36.44	36.11
Triphone + Δ + Δ	36.47	36.98	36.77
+ MLLT + LDA	27.24	27.71	26.80
+ SAT(fMLLR)	20.82	21.90	**20.60**

4.2. System combination

Using lattices from the best WERs (see Table 7), we combine several systems for decoding. Kaldi supports system combination based on minimum Bayes risk (MBR) [23] decoding. It combines lattices from a number of systems and produces sequences that have least expected losses. Our

combination strategies and results are described in Table 8 where we ran 2 and 3-system combination. Overall, the results are better compared to results through single lattice decoding. For the 2-system combination, Hybrid G2P + Iban G2P gave less improvement than Hybrid G2P + Malay G2P. While Malay G2P + Iban G2P had an average between the previous two. Interestingly, an addition of Iban G2P to the Hybrid G2P + Malay G2P combination gave the best result among others.

Table 8. System combination and WERs

Combination	%WER
Hybrid G2P + Iban G2P	19.83
Malay G2P + Iban G2P	19.76
Hybrid G2P + Malay G2P	19.55
Hybrid G2P + Malay G2P + Iban G2P	19.22

5. RESULTS ANALYSIS

5.1. Weak correlation between ASR and G2P performances

Figure 1 presents a graph plot of ASR and G2P results which we have previously obtained. The ASR values are the best results from each decoders while G2P results are estimated values taken from Table 4. From this graph, we can observe the performance of the Hybrid system is the best among the other two systems where G2P and ASR accuracies are 29.8% and 20.6% WERs, respectively. However, the correlation between ASR and G2P performances is rather weak. The bad news is that all our efforts to improve our Iban lexicon (with the hybrid approach) did not have a strong effect on ASR. The good news is that using G2P of a similar well-resourced language (such as Malay) seems to be a good starting point to generate pronunciations and build an ASR system for a very under-resourced language (such as Iban).

5.2. Confusion pairs

For final evaluation, we conducted a confusion analysis to observe words that were wrongly recognized (substitution). To perform this, we obtained all confusion pairs (generated by NIST toolkit [24]) by utilizing outputs from best results as shown in Table 7 and Table 8 as well as the reference transcript. Table 9 presents the top ten most frequent confusion pairs. Words on the left are words in the reference while words on the right are the outputs. The first four columns are pairs from the reference and outputs of the single systems and 3-system (Hybrid G2P + Malay G2P + Iban G2P). Meanwhile the last column shows pairs of outputs from Malay G2P and Iban G2P systems.

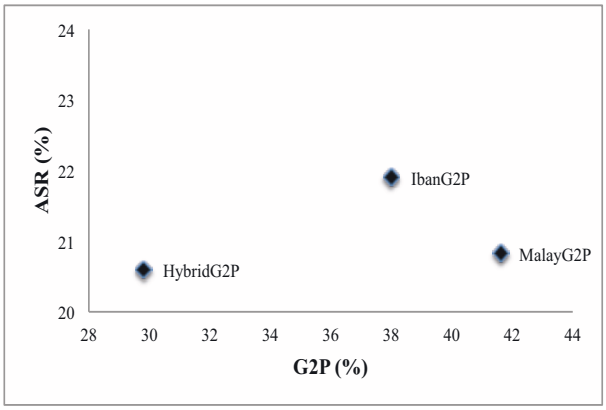

Fig. 1. Iban ASR vs G2P based on WER results

Overall, there are normalization issues and morphological errors that can be observed from this table. An example of a normalization problem that we can point out is, the word *rakyat* (people) is a Malay word and the system recognized *rayat*, which is actually correct for Iban. The mistake exists in the reference which results penalizing recognition performance. A possible reason is that transcribers could have been influenced by Malay spellings when creating the speech transcript. Other examples are such as *urang* and *orang* (person), *serta* and *sereta* (as well as / join), mohamad and mohd, *penerbai* and *penerebai* (airline), *agensi* and *ijinsi* (agency) and, *ka* and *ke*. For the case of *ti* and *ke*, both are Iban words where the former is a conjunction and the latter is an adjective. Though orthographically different, both are synonyms and used frequently to describe things or people [25]. As for *dato* and *datuk*, these words are pronunced as /dato?/ . These are titles awarded by the head of states or sultan; *Dato* (here apostrophe is neglected, original is *Dato'*) or *Datuk* is placed before a person's name. Some pairs that have morphological problems are such as *ka* with *madahka* (also with *bejalaika* or *ngambika*), waifm and fm, as well as *sehari* and *tu* (can originate from the word *seharitu* / *saritu*; found two versions; pronounce as /saritu?/; means today). In Iban, *ka* is a suffix that forms transitive verbs just like the suffix *kan* in Malay. Apparently, this suffix is separated frequently from the root words in the Iban text and speech transcript.

6. CONCLUSIONS AND PERSPECTIVES

The paper demonstrates our effort in obtaining an ASR for Iban, the first system for this language from Borneo island. The close relationship between Malay and Iban, where both belong to the same language branch, motivated us to propose a bootstrapping strategy to generate a phonetic transcript for Iban from a Malay one. The generated sequences were manually post-edited and the post edited version was later

Table 9. Top ten confusion pairs from Hybrid, Malay, Iban systems and system combination

Hybrid	Malay	Iban	Combine (H+M+I)
rakyat => rayat	**rakyat => rayat**	**rakyat => rayat**	**rakyat => rayat**
ka => ke	**ka => ke**	**ari => hari**	**ka => ke**
ti => ke	**ari => hari**	**ka => ke**	**ari => hari**
ari => hari	**serta => sereta**	**serta => sereta**	**ti => ke**
urang => orang	**ti => ke**	**ti => ke**	**serta => sereta**
serta => sereta	**urang => orang**	**urang => orang**	**urang => orang**
mohamad => mohd	**datuk => dato**	**ke => ka**	**ke => ka**
ka => madahka	ka =>madahka	**mohamad => mohd**	**mohamad => mohd**
ke => bejalaika	ke => bejalaika	**agensi => ijinsi**	**agensi => ijinsi**
antara => entara	**mohamad => mohd**	ka => madahka	ka => madahka

Bold : normalization problem; Normal : morphological problem

used for Iban G2P training. Our G2P evaluation results prompted us to phonetize 37K Iban words using two G2Ps, Malay (68K) and Iban (1K). As a result, we have a mix of Malay and Iban pronunciations in this Hybrid G2P. In addition, we developed two other lexicons, each of them was produced by either Iban or Malay G2P.

We built three Iban ASRs that use three different pronunciation dictionaries; Malay, Iban and mix (Hybrid). To conduct this investigation, the acoustic material consisted of almost 7 hours of training and one hour of test material. Various acoustic modeling techniques were employed to test the systems. For the language model, we utilized news text for training. A trigram language model was trained on 2.08M words and we obtained a perplexity of 158 and 2.3% OOV rate after an evaluation using the speech transcript.

Our best results for Iban ASR (with different lexicon) were as follows: Malay G2P (20.82%), Iban G2P (21.90%) and Hybrid G2P (20.60%). These results were produced after fMLLR adaptation. In this paper, we also reported some analysis such as correlation between ASR and G2P and confusion pairs. Furthermore, we attempted system combination to decode and our best result was 19.22% WER for Hybrid G2P + Malay G2P + Iban G2P combination.

Through our experiments, we found that using Malay G2P dictionary alone can help our system to achieve a very acceptable ASR result. Interestingly, training a dedicated Iban G2P system (from a small training data) did not help much (a difference of 1%). Finally, a hybrid version of both Malay and Iban pronunciation lexicons was able to improve the ASR performance slightly.

Following these results, we plan to further work on several issues pertaining to ASR and our data. We would like propose solutions to solve the normalization problems found in the speech transcript and text data. This is important as we believe that by reducing orthography mistakes in these texts will be able to reduce the errors in the outputs. Another point that we would like to work on is to develop an ASR that is trained on subspace Gaussian mixture models (SGMMs), as a solution to further improve our baseline results.

7. REFERENCES

[1] Josef R. Novak, Nobuaki Minematsu, and Keikichi Hirose, "Evaluations of an open source wfst-based phoneticezer," PDF, General Talk No. 452, The Institute of Electronics, Information and Communication Engineers, 2011.

[2] Maximilian Bisani and Hermann Ney, "Joint-sequence models for grapheme-to-phoneme conversion," *Speech Communication*, vol. 50, no. 5, pp. 434–451, 2008.

[3] Sittichai Jiampojamarn and Grzegorz Kondrak, "Online discriminative training for grapheme-to-phoneme conversion," in *Proc. INTERSPEECH*, 2009, pp. 1303–1306.

[4] Marelie Davel and Olga Martirosian, "Pronunciation dictionary development in resource-scarce envionments," in *Proc. INTERSPEECH*, 2009, pp. 2851–2854.

[5] Sameer R. Maskey, Alan W Black, and Laura M. Tomokiyo, "Bootstrapping phonetic lexicons for language," in *Proc. INTERSPEECH*, 2004, pp. 69–72.

[6] Laurent Besacier, Etienne Barnard, Alexey Karpov, and Tanja Schultz, "Automatic speech recognition for under-resourced languages : A survey," *Speech Communication Journal*, vol. 56, pp. 85–100, January 2014.

[7] W. Heeringa and F. de Wet, "The origin of afrikaans pronunciation: a comparison to west germanic languages and dutch dialects," in *Proc. Conference of the Pattern Recognition Association of South Africa*, 2008, pp. 159–164.

[8] M. P. Lewis, Gary F. Simons, and Charles D. Fennig, *Ethnologue : Languages of the world, SIL International*, 2013. Available at : http://www.ethnologue.com.

[9] Matthew S. Dryer and Martin Haspelmath, Eds., *WALS Online*, Max Planck Institute for Evolutionary Anthropology, Leipzig, 2013. Available at : http://wals.info/.

[10] Adelaar and Alexander, *The Austronesian Languages of Asia and Madagascar: A Historical Perspective, The Austronesian Languages of Asia and Madagascar*, Routledge Language Family Series, London, 2005.

[11] Sarawak Education-Department, *Sistem Jaku Iban di Sekula*, Sarawak, Malaysia, 1st edition, 2007.

[12] Asmah Haji Omar, *Phonology*, Dewan Bahasa dan Pustaka, Kuala Lumpur, Malaysia, 1981.

[13] Andreas Stolcke, "Srilm - an extensible language modeling toolkit," in *Proc. of the 7th International Conference on Spoken Language Processing (ICSLP 2002)*, 2002, pp. 901–904.

[14] C. Barras, E. Geoffrois, Z. Wu, and M. Liberman, "Transcriber:development and use of a tool for assisting speech corpora production," in *Proc. Speech Communication special issue on Speech Annotation and Corpus Tools*. 2000, vol. 33, available at : trans.sourceforge.net/en/publi.php.

[15] C. Barras, E. Geoffrois, Z. Wu, and M. Liberman, "Transcriber: a free tool for segmenting, labeling and transcribing speech," in *Proc. First International Conference on Language Resources and Evaluation (LREC)*, 1998, pp. 1371–1376.

[16] Tien-Ping Tan, H. Li, E. K. Tang, X. Xiao, and E. S. Chng, "Mass: a malay language lvsr corpus resource," in *Proc. Oriental COCOSDA International Conference 2009*, 2009, pp. 26–30.

[17] Josef R. Novak, "Phonetisaurus: A wfst-driven phoneticizer. available at : https://code.google.com/p/phonetisaurus," 2012.

[18] Sarah Samson Juan and Laurent Besacier, "Fast bootstrapping of grapheme to phoneme system for under-resourced languages - application to the iban language," in *Proc. 4th Workshop on South and Southeast Asian Natural Language Processing 2013*, Nagoya, Japan, October 2013.

[19] V. I. Levenshtein, "Binary codes capable of correcting deletions, insertions and reversals.," in *Soviet Physics-Doklady*, 1966, vol. 10, pp. 707–710.

[20] Daniel Povey, Arnab Ghoshal, Gilles Boulianne, Lukas Burget, Ondrej Glembek Nagendra Goel Mirko Hannemann Petr Schwarz, Jan Silovsky, Georg Stemmer, and Karel Vesely, "The kaldi speech recognition toolkit," in *IEEE 2011 Workshop on Automatic Speech Recognition and Understanding*, IEEE Signal Processing Society, Ed., December 2011, vol. IEEE Catalog No. : CFP11SRW-USB.

[21] R. A. Gopinath, "Maximum likelihood modeling with gaussian distributions for classification," in *Proc. of ICASSP*, 1998, pp. 661–664.

[22] M. Gales, "Maximum likelihood linear transformations for hmm-based speech recognition," in *Computer Science and Language*, 1998, vol. 12, pp. 75–98.

[23] V. Goel, S. Kumar, and W. Byrne, "Segmental minimum bayes-risk decoding for automatic speech recognition," in *Proc. IEEE Transactions on Speech and Audio Processing*, 2003.

[24] NIST, "Speech recognition scoring toolkit (sctk). available at : http://www.nist.gov/speech/tools/," 2010.

[25] Janang Ensiring, Jantan Umbat, and Robert Menua Saleh, *Bup Sereba Reti Jaku Iban*, The Tun Jugah Foundation, Sarawak, Malaysia, first edition, 2011.