HAL
archives-ouvertes.fr

USING MALAY RESOURCES TO BOOTSTRAP ASR FOR A VERY UNDER-RESOURCED LANGUAGE: IBAN

Sarah Samson Juan, Laurent Besacier, Solange Rossato

▶ **To cite this version:**

Sarah Samson Juan, Laurent Besacier, Solange Rossato. USING MALAY RESOURCES TO BOOTSTRAP ASR FOR A VERY UNDER-RESOURCED LANGUAGE: IBAN. SLTU 2014, May 2014, Saint-Petersbourg, Russia. hal-01002920

HAL Id: hal-01002920
https://hal.inria.fr/hal-01002920

Submitted on 7 Jun 2014

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

USING MALAY RESOURCES TO BOOTSTRAP 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.

Index Terms— under-resourced languages, speech recognition, Iban language, Malay language, bootstrapping, Kaldi

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. One way to undertake this problem is by listing all pronunciations in the lexicon (G2P conversion). However, this requires time to manually list thousands of pronunciations. Currently, there are data driven techniques to help train G2P systems such as (joint-sequence model papers) 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 South African and Nepali ([1], [2]). This semi-supervised method predicts additional entries of a dictionary through a pronunciation model and the outputs are then verified by a native speaker / 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 [3].

Our paper introduces a feasible approach that is suitable for languages in the same family (same orthography, phonetically similar). We propose to use Malay pronunciations to produce a base transcript for Iban and then, post-edit the outputs. This idea comes from the fact that Malay and Iban is closely related and both belong to the same language family. 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 [4] conducted 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 phonemes, we train an Iban G2P system to phonetize other entries. Both Malay and Iban G2Ps are evaluated, whereafter, results suggesting a (supervised???) 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 available resources for Iban ASR including 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 give perspectives.

2. THE IBAN LANGUAGE

Iban is a member of the Malayo-Polynesian language family, under the Ibanic group. The language belongs to the same family as Malay, where, the latter is under to the Malayic group [5]. With over 600 thousand Iban speakers, it is mostly spoken in Sarawak, Kalimantan and Brunei. In the course of modernization, there are also Iban speakers found in

the Peninsular Malaysia, for example, in Johor and Kuala Lumpur. Alongside learning Malay, Iban has been taught in schools from primary to secondary level as a nonobligatory subject since the early 90s. At several universities, basic Iban courses are offered to undergraduate students in Malaysia. The Iban system is influenced by the Malay system in terms of phonology, morphology and syntax [6]. According to [6], a guidebook on Iban, there are many words that belong to both languages (same surface forms) and borrowed words from Malay. Example of same surface forms are listed in Table 1 together with their phoneme sequences.

Table 1. *Malay/Iban examples with phonetic representations*

No.	Words	Iban	Malay
1	ke	/kə/	/kə/
2	nya	/ɲaʔ/	/ɲə/ or / ɲa/
3	iya	/ija/	/ija/
4	bilik	/biliəʔ/	/bileʔ/
5	dua	/duwa/	/duwə/ or /duwa/
6	sida	/sidaʔ/	/sida/
7	puluh	/puluəh/	/puloh/
8	raban	/raban/	/raban/
9	lalu	/lalu/	/lalu/
10	orang	/uraŋ/	/oraŋ/

In 1981, [7] published the first 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/, /dʒ /, /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 borrowed consonants (from Malay) such as /f/, /v/, /θ/, /z/, /x/, /ɣ/, /ð/, /ʃ/. As shown in Table 1, Iban and Malay orthographies are Latin based where both languages use 26 English alphabets. Moreover, Iban is not a tonal language like Malay. The obvious differences between Malay and Iban are the appearance of vowel clusters or transition of two vowels within two consonants and more /ʔ/ sounds for words ending with vowels.

3. IBAN RESOURCES

3.1. Text data

We utilized Iban electronic texts as our data. News data was collected from a news website [1] 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.

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

3.2. Language model

Using this text data, we built a trigram Iban language model with modified Kneser-Ney discounting. SRILM Toolkit was used to obtain the model and later, measured the model's perplexity on Iban speech transcript. The evaluation gave us a perplexity of 162 and 2.3% OOV rate.

3.3. Speech corpus and transcript

We have eight hours of news data with 16khz sampling rate. The data was provided by the Radio Televisyen Malaysia (RTM), a local radio and television station, through one of its channel called 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.

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 and test sets for the experiment

Set	Speakers	Gender (M:F)	Sentences	(mins)
Evaluation	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 ([8], [9]), 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.

3.4. Pronunciation dictionary

3.4.1. Obtaining the Iban G2P system

In the Iban text data, we found 37K unique words. The list has Malay (23%) and English (19%) words, a verdict we made after conducting a comparison study with Malay (list from [10]) and English (CMU version for Sphinx) vocabularies. Following that, were intrigued to know the pronunciation distances between Malay and Iban, particularly the same surface forms (hereafter, we address as Malay-Iban). To implement this, we applied Levenshtein method to calculate the distances, following a similar study conducted by [4] 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 have closer pronunciations to Afrikaans compared to the other two languages based on the phonetic transcripts.

We tested on 100 most frequent Malay-Iban words in the Iban text. First, we obtained a Malay pronunciation lexicon from [10]. Tan et al. in [10] developed a 76K pronunciation lexicon for their Malay speech recognition that has a baseline ASR result of 14.6% WER. Using their lexicon, we trained a Malay grapheme to phoneme (G2P) as a base G2P system using Phonetisaurus, an open source tool based on Weighted Finite States Transducers ([11], [12]). 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 [13] for more details).

Next, phoneme sequences for Malay were generated and then post-edited to match Iban pronunciations. The latter part was done by an Iban native speaker[2]. At this stage, we limit to Malay phoneme set. 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 47 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

Note: IM for Malay-Iban words and I for pure Iban words. [13]

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 previous) and our results in [13] showed that Malay G2P can phonetize Malay-Iban better than pure Iban, whereas Iban G2P works better for pure Iban (see Table 4).

3.4.2. Obtaining pronunciations for the whole lexicon

After the analyzing the Malay and Iban G2P performances, 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 [10]'s work for Malay recognizer. Using this proposed strategy, we have 37K pronunciations including 1K of Iban G2P data. The pronunciation lexicon is estimated to have 8.1% PER and 29.4% WER on 2K random outputs.

3.4.3. Analyzing pronunciations

Besides having a mix of Malay and Iban pronunciations in the dictionary (later address as Hybrid G2P), we also 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 compare two dictionaries and our findings are presented as 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** in both $\mathbf{A}$ and $\mathbf{B}$ or can be described as, $C_{\mathbf{AB}} = \{(x_i, y_i) \mid x_i \in \mathbf{A}, y_i \in \mathbf{B}, x_i \neq y_i, \forall i \in [1, N]\}$where $\mathbf{A}$ and $\mathbf{B}$ are two pronunciation lists and N is the total number of pronunciations. From Table 5, 67% of Malay G2P dictionary is different than Iban G2P, the highest number of differences compared to the other two comparison pairs. Meanwhile, the hybrid version is closer to Iban G2P with 29% error.

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

$C_{\mathbf{AB}}$	No. of 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 did the words with different pronunciations (elements

[2]the first author of this paper

of $C_{\mathbf{AB}}$) belong to. As mentioned in Section 3.4.1, 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 Malay G2P to Iban G2P, majority of the differences belong to pure Iban and the same can be said when we compared Malay G2P to Hybrid G2P. As English words were phonetized by Iban G2P system, therefore these pronunciations are different with Malay G2P and Hybrid G2P (5,605) and no differences between Malay G2P and Hybrid G2P for these. Note that there are Iban G2P training data available in Hybrid G2P ($S3$), which means there are post-edited pronunciations. Hence, there are some differences for Malay and pure Iban words when Hybrid G2P was compared to Malay G2P and Iban G2P, respectively.

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 [14] for Iban, an open source toolkit based on FST. 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. For the training, we explored several techniques offered by Kaldi. For this study, 13 MFCCs were extracted and GMM models were employed for monophone and triphone trainings. For triphone, we use 4,200 context-dependent states and 40,000 Gaussians. We also implemented delta delta coefficients on the MFCCs, LDA transformation together with MLLT, and, speaker adaptation with and without feature spaced MLLR. Moreover, decoding was launched with language model scales of 5.0 to 20.0 thus, resulting 16 WERs per decoding.

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	25.86	27.02	25.82
+ fMLLR	20.82	21.90	20.60

4.2. System combination

Using lattices from the best WERs(see Table 7), we combine the N systems for decoding. Kaldi supports system combination based on Minimum Bayes Risk (MBR) decoding. It combines lattices from several systems and produces sequences that have least expected losses. Our combination strategies and results are described as in Table 8 where we ran this for 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. Correlation between ASR and G2P

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.

5.2. Inter and Intra-system : Match HYPs

We analyzed hypotheses (HYP) or outputs generated by the Iban ASRs in order to measure the diversity of sequences.

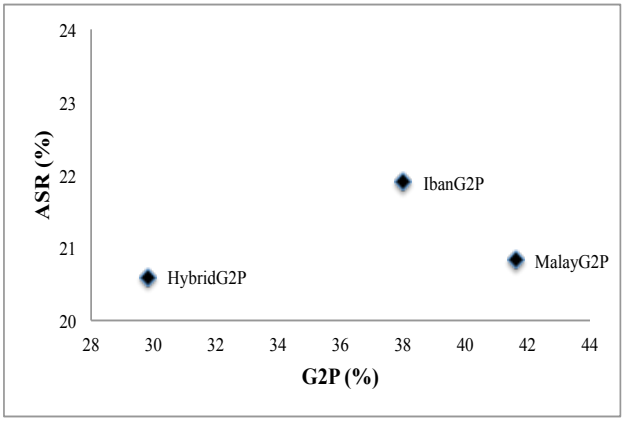

Fig. 1. Iban ASR vs G2P based on WER results

As different language model weights were applied during the decoding, we have a total of 16 HYP transcripts. From these scripts, we acquired three from each system where all of them are the top three best results. Subsequently, we made an intra-system comparison and results are shown in Table 9. Based on the values we obtained, one HYP is between 78 to 90.3 percent equivalent to another HYP. Conclusively, the best HYPs (1st) are close to the second best HYPs (2nd) and third best HYPs (3rd) but 2nd and 3rd HYPs are 10% more different than the former HYP pairs.

Table 9. Matching HYP pairs (%); taken from top three best ASR results from each system

ASR	HYP		
	1st-2nd	2nd-3rd	1st-3rd
Malay G2P	90.2	79.7	87.7
Iban G2P	87.7	79.3	90.3
Hybrid G2P	89.2	78	86.9

The following analysis was an inter-system comparison to find common HYP pairs across two or three ASR systems. This time, the results are less diverse compared to intra-system. It is well noted that Hybrid G2P have higher number of common sequences with Malay G2P outputs than Iban G2P outputs (see Table 10). These results are in parallel with system combination outputs as shown in Table 8; a combination of Hybrid G2P with Malay G2P system gave better result compared to Hybrid G2P combine with Iban G2P. When Hybrid G2P was compared to Malay G2P and then to Iban G2P, we found less than 12% pairs matched.

5.3. 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[15]) based on outputs from the best results as shown in Table 7 and Table 8 as well as the reference transcript. Table 11 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.

Table 10. Matching HYP pairs (%) found across two and three ASR systems.

ASR	Malay G2P	Iban G2P
Hybrid G2P	22.4	20.1
Malay G2P	-	17.5
Hybrid G2P	11.8	

Overall, there are normalization issues and morphological errors can be observed from this table. An example of a normalization problem that we can see, 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 [16]. As for *dato* and *datuk*, these words have same pronunciation /dato?/ and they 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* and *madahka*, *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). *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 Polynesian language. The close relationship between Malay and Iban, where both belongs to the same language family, 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 used for Iban G2P training. Our G2P evaluation results prompted us

Table 11. Top ten confusion pairs from Hybrid, Malay, Iban systems and system combination

Hybrid	Malay	Iban	Combine (H+M+I)	Malay vs. Iban
rakyat => rayat	rakyat => rayat	rakyat => rayat	rakyat => rayat	ke => ka
ka => ke	ka => ke	ari => hari	ka => ke	dato => datuk
ti => ke	ari => hari	ka => ke	ari => hari	skim => sekim
ari => hari	serta => sereta	serta => sereta	ti => ke	ke => ti
urang => orang	ti => ke	ti => ke	serta => sereta	seri => sri
serta => sereta	urang => orang	urang => orang	urang => orang	hari => ari
mohamad => mohd	datuk => dato	ke => ka	ke => ka	sehari => tu
ka => madahka	ka =>madahka	mohamad => mohd	mohamad => mohd	ngambika => ngambi
ke => bejalaika	ke => bejalaika	agensi => ijinsi	agensi => ijinsi	penerbai => penerebai
antara => entara	mohamad => mohd	ka => madahka	ka => madahka	waifm => fm

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 162 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 feature spaced MLLR adaptation was applied. In this paper, we also reported other results such as intra and intersystem hypothesis analysis, correlation between ASR and G2P, confusion pairs analysis and others. Furthermore, we attempted system combination to decode and our best results 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 favourable ASR result. Interestingly, the Iban G2P system that has a pronunciation lexicon generated solely by a small training Iban data, was able to achieve almost the same result as Malay G2P (a difference of 1%). However, it is well noted that a hybrid version of the pronunciation lexicon was able to improve the ASR results.

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 GMM models, as a solution to further improve our baseline results.

7. REFERENCES

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

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

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

[4] 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.

[5] M. P. Lewis, Gary F. Simons, and Charles D. Fennig, "Ethnologue : Languages of the world, sil international. available at : http://www.ethnologue.com," 2013.

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

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

[8] 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.

[9] 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.

[10] 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.

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

[12] 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.

[13] 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.

[14] 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.

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

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