Using Resources from a Closely-related Language to Develop ASR for a Very Under-resourced Language: A Case Study for Iban

Sarah Samson, Laurent Besacier, Benjamin Lecouteux, Mohamed Dyab

Using Resources from a Closely-related Language to Develop ASR for a Very Under-resourced Language: A Case Study for Iban

Sarah Samson Juan[1,2], *Laurent Besacier*[2], *Benjamin Lecouteux*[2], *Mohamed Dyab*[2]

[1]Faculty of Computer Science and Information Technology, UNIMAS, Malaysia
[2]Grenoble Informatics Laboratory (LIG), Univ. Grenoble Alpes , Grenoble, France
[1]sjsflora@fit.unimas.my, [2]{laurent.besacier, benjamin.lecouteux}@imag.fr,
[2]muhamed.dyab@gmail.com

Abstract

This paper presents our strategies for developing an automatic speech recognition system for Iban, an under-resourced language. We faced several challenges such as no pronunciation dictionary and lack of training material for building acoustic models. To overcome these problems, we proposed approaches which exploit resources from a closely-related language (Malay). We developed a semi-supervised method for building the pronunciation dictionary and applied cross-lingual strategies for improving acoustic models trained with very limited training data. Both approaches displayed very encouraging results, which show that data from a closely-related language, if available, can be exploited to build ASR for a new language. In the final part of the paper, we present a zero-shot ASR using Malay resources that can be used as an alternative method for transcribing Iban speech.

Index Terms: speech recognition, low resource languages, cross-lingual training, zero-shot ASR

1. Introduction

Building automatic speech recognition system (ASR) for under-resourced languages has many constraints such as: lack of transcribed speech, few speaker diversity for acoustic modelling, no pronunciation dictionary and few written materials to build language models [1]. It requires a considerable amount of time to gather resources, if there are any, thus under-resourced languages generally have a very limited amount of data to train current GMM and DNN based systems. Past studies have shown that cross lingual or multilingual acoustic models can help to boost the performance of language-specific systems by providing universal phone units that cover several spoken languages (e.g. [2],[3],[4],[5]). However, mapping source phone units to target units can be tricky, especially for very under-resourced languages that are poorly described.

Lately, Subspace Gaussian Mixture Models (SGMM) ([6], [7]) have shown to be very promising for ASR in limited training conditions ([8], [9]). In this improved technique for HMM/GMM system, the acoustic units are all derived from a common GMM called Universal Background Model (UBM). The globally shared parameters of this UBM do not need the knowledge about the phone units used in the source language(s). Without this constraint of source-target mapping of acoustic units, the UBM can be easily used in cross-lingual or multilingual settings. Furthermore, UBM is trained on data that has many speakers which also help to increase speaker diversity in the acoustic space. In the mean time, Deep Neural Networks (DNNs) have been increasingly employed for building efficient ASR systems. HMM/DNN hybrid systems clearly outperform HMM/(S)GMM systems for many ASR tasks [10] which include dealing with low-resource systems ([11], [12], [13]). Several studies have shown that multilingual DNNs can be achieved by utilizing multilingual data for conducting unsupervised RBM pretraining [14] or training the whole network ([13], [12], [15]).

Most of the cross-lingual works cited above used one or several "source" language to help the design of "target" language ASR. However, the choice of the source language(s) was not always legitimate while we believe that the use of a closely-related (well resourced) language is the best option in most cases. So, this paper tries to answer to the following question: is there a clear benefit when using resources from closely-related language for developing ASR for a very under-resourced language? We evaluate this not only for cross-lingual acoustic modelling but also for semi-supervised pronunciation lexicon design. Our target language is Iban spoken in Sarawak (part of Malaysia on Borneo island) and we systematically compare the use of closely-related language (Malay) resources with the use of non closely-related language (English) resources for Iban ASR. The rest of the paper is organized as follows. In Section 2, we describe the starting point of our work: Iban data available. Section 3 presents a semi-supervised approach for building Iban pronunciation dictionary using out-of-language resources. In Section 4, we compare cross-lingual acoustic modelling approaches using both SGMM (less efficient than DNNs but more compact for embedded applications) and DNN (state-of-the-art) frameworks (with Malay or English as the "source" languages). Section 5 also investigates a zero-shot ASR training procedure using ASR in a closely-related language to generate speech transcripts. Last but not least, Section 6 concludes this paper.

2. Starting point : Iban data available

Iban is a language spoken in Borneo, mainly in Sarawak (Malaysia), Kalimantan and Brunei. The target language is closely-related to Malay, a local dominant language in Malaysia. Both languages are part of the Austronesian language family, in the branch of Malayo-Polynesian languages [16]. In [17], we have presented the relationship between the two languages based on several studies and our observations. Our Iban corpora contain speeches for acoustic modelling and texts for language modelling (no pronunciation dictionary was available initially). The speech corpus was obtained from *Radio Televisyen Malaysia Berhad*, a local radio and television station in Malaysia. We have almost eight hours of clean speech, spoken by 23 speakers (17 speakers or 6h48 for train

- 6 speakers or 1h11 for test). After manually transcribing the speech data, we have acquired more than 3K sentences. The text corpus contains 2M words, obtained from online news articles[1]. We performed text normalization to remove punctuations, HTML tags, transcribing numbers to letters and abbreviations to full terms. In total, there are 37K unique words available in the text.

3. Semi-supervised approach for building Iban pronunciation dictionary

Creating a pronunciation dictionary from scratch, especially for a new language is typically time-consuming and requires an expert (if available) of the target language to complete the task. Bootstrapping grapheme-to-phoneme (G2P) system was introduced in [18] to reduce effort of producing phonetic transcriptions manually. The semi-supervised method was used to produce a pronunciation dictionary for Nepali and English. It requires a small transcript with words and pronunciations in a target language, which is usually prepared by a native speaker or a linguist. The transcript then becomes the seed data for building G2P rules. Then, the pronunciation model is used to predict new entries in the vocabulary and post-editing is conducted after, if needed. The post-edited data is used to update the model and this process can be repeated for obtaining better estimates. The method was also used by [19] for German and Afrikaans [20] languages. This idea motivates us to investigate bootstrapping G2P for Iban. The first question is whether we can use data of a closely-related language to quickly build a system for the target language.

3.1. Semi-supervised lexicon design

We employed the following strategy for acquiring a pronunciation dictionary for Iban. The description is a brief one, since preliminary experiments with this approach were presented in [21]. First, we obtained a Malay pronunciation dictionary which was used for building a Malay ASR (MASS corpus [22]). We trained a **Malay G2P** on Phonetisaurus ([23], [24]) G2P toolkit using 68K pronunciations and 34 phonemes. Then, the G2P system was used to phonetize 1K Iban words for obtaining a base pronunciation transcript. The outputs were subsequently corrected by a native speaker[2] and later we used the post-edited pronunciations to build an **Iban G2P**. We also propose a lexicon based on the following approach (later called **Hybrid G2P**): the Malay G2P phonetizes all common (same surface forms) Malay-Iban words while the Iban G2P phonetizes only pure Iban words (both G2P systems use the same number of phonemes). In addition, we use two systems to phonetize Iban lexicon; a grapheme-based system (**Grapheme**) which we obtained from authors in [25] and an **English G2P** built using English CMU dictionary - a demo system for Phonetisaurus. The latter allows us to compare the use of a non-closely related language resource (English) with the use of a closely-related one (Malay).

3.2. Evaluation of pronunciation dictionaries for an ASR task

We used Kaldi speech recognition toolkit [26] for building our ASR systems. For this preliminary evaluation of the pronunciation lexicon, we limited our experiments to a HMM/GMM acoustic model. The acoustic models were built using 13 MFCCs and Gaussian mixture models (GMM) on 7h training data. We trained triphone models by employing 2,998 context-dependent states and 40K Gaussians. Besides that, we implemented delta delta coefficients on the MFCCs, linear discriminant analysis (LDA) transformation and maximum likelihood transform (MLLT) [27], as well as speaker adaptation based on feature-space maximum likelihood linear regression (fMLLR) [28]. Each Iban system used either one of the pronunciation dictionaries available for the training and decoding stages. In the decoding process, we also used a trigram language model with modified Kneser-Ney discounting applied. The language model was trained on our 2M words Iban news data using SRILM [29].

Table 1: *Impact of Pronunciation Dictionaries on Iban (HMM/GMM) ASR performance (WER %) on our test set (1h)*

Pronunciation Dictionary	Triphone model	
	no spkr adapt	spkr adapt
Grapheme	32.9	20.5
English G2P	39.2	22.9
Malay G2P	35.7	19.8
Iban G2P	36.2	20.9
Hybrid G2P	36.0	**19.7**

Table 1 shows the evaluation results of Iban ASR for different pronunciation dictionaries and acoustic models. Note that the results are slightly different than the ones shown in [21] because we have updated our test transcription data by solving several text normalization issues after the previous paper was published. The systems performed better when developed using speaker adaptive training approach. We achieved 12-17% absolute improvement on the word error rates (WERs) over the non speaker adapted systems results. In terms of pronunciation dictionary, Hybrid G2P provided the best result for this training approach. Interestingly, this result is closely followed by Malay G2P. Thus, the ASR results proved that Malay data can be used as a starting point for building Iban pronunciation dictionary from scratch. The results are significantly better with Malay than with English which is not closely related to Iban. Besides that, we also observe that using grapheme-based approach can be also a good option in our case. In the following sections, we use the best (Hybrid) G2P for all our ASR experiments.

4. Cross-lingual approaches for building Iban acoustic models

4.1. Brief background on cross-lingual acoustic modelling for low-resource conditions

4.1.1. Subspace Gaussian Mixture Model

The GMM and SGMM acoustic models are similar since each emission probability of each HMM state is modelled with a Gaussian mixture model. However, in the SGMM approach [7], the Gaussian means and mixture component weights are generated from the phonetic and speaker subspaces along with a set of weight projections. These globally shared parameters are common across all states. [8] and [9] presented cross-lingual and multilingual work using SGMM for improving ASR with very limited training data. In both studies, the authors carried out the cross-lingual approach by employing UBM trained on source language data, either monolingual or multilingual data, in SGMM training of their target language. Applying this technique improved the ASR performance of their low-resource system.

Inspired by the studies mentioned above, we try to find what is the best source language (or source languages combination) to use in the cross-lingual SGMM method for Iban ASR. As in the previous section: Malay is the closely-related language

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

[2] The first author of this paper

while English is the non closely-related one.

4.1.2. Deep Neural Network

Deep Neural Network (DNN) for ASR is a feed-forward neural network with hidden layers. Optimizing hidden layers can be done by pretraining using Restricted Boltzmann Machines (RBM) [30]. The generative pretraining strategy builds stacks of RBMs corresponding to the number of desired hidden layers and provides better starting point (weights) for DNN fine-tuning through backpropagation algorithm. Pretraining a DNN can be carried out in a unsupervised manner because it does not involve specific knowledge (labels, phone set) of a target language[3]. Only the softmax layer is sensitive to the target language. It is added on top of the hidden layers during fine-tuning and its output corresponds to the HMM states of the target language. As shown in [14], using untranscribed data for RBM pretraining as a multilingual strategy has little effect on improving monolingual ASR performance. The *transfer learning* [15] approach has shown large recognition accuracy improvements. The technique involves removing the top layer of a multilingual DNN and fine-tuning the hidden layers to a specific language. This cross-lingual approach has been applied in several studies for improving low-resource systems. Such studies can be found in [11], [12] and [13]. In this paper, we investigate an approach for obtaining a language-specific DNN for Iban ASR. The method is applied for observing impact of using hidden layers from Malay to train DNN for Iban. Additionally, we employ a strategy to transfer speaker adapted DNNs.

4.2. Training ASR (SGMM, DNN) on Iban data only

We trained SGMM and DNN acoustic models on two conditions of (transcribed) training speech: 1h and 7h Iban data. The 1h data was randomly picked from the 7h training data. The 7h HMM/GMM system is already described in the previous section. For the 1h system, we trained a triphone model (39 MFCC with deltas and deltas deltas) using 664 context-dependent states and 5K Gaussians. We did not include speaker adaptive training at this point, because we want to observe only the cross-lingual effect in the experiments described in this section. The UBM was trained on 7h of untranscribed (training) data for initializing SGMMs of both systems, using 600 UBM Gaussians and the phonetic subspace dimension was set to 40. Then, the SGMMs were derived from this UBM with 805 substates in the 1h system and 10K substates in the 7h system. To build DNNs, we trained the network using state-level minimum Bayes risk [31] (sMBR) and the network has seven layers, each of the six hidden layers has 1024 hidden units. The network was trained from 11 consecutive frames (5 preceding and 5 following frames) of the same MFCCs as in the GMM systems. Furthermore, same HMM states were used as targets of the DNN. The initial weights for the network were obtained using Restricted Boltzmann Machines (RBMs) that resulted in a deep belief network with 6 stacks of RBMs. Fine tuning was done using Stochastic Gradient Descent with per-utterance updates, and learning rate 0.00001 kept constant for 4 epochs.

The ASR results for monolingual SGMM and DNN are presented in Table 2a. Both modelling techniques provided better ASR performance than GMM for the two systems with different amount of training data. For the 1h system, with SGMM and DNN we achieved 2.5% and 13.4% absolute improvements on the WER, respectively. On the other hand, both approaches resulted almost equal performance in the 7h system. We achieved 17.1% and 17.6% reductions in WER for the SGMM and DNN systems, respectively.

[3]In that sense, RBM pretraining (for DNN) and UBM training (for SGMM) are both unsupervised methods to get an initial representation of the acoustic space before modelling the speech units

4.3. Out-of-language resources for Iban ASR

Our research question is whether using data from Malay in cross-lingual acoustic modelling for Iban can improve the performance of the low-resource ASR. To answer this question, we employed two out-of-language databases in our experiments. We used Malay speech corpus as data from a closely-related language and English corpus as data from a *non* closely-related language. Given these two scenarios, we hope to evaluate Malay and English performances for cross-lingual SGMM and language-specific DNN. The MASS corpus [22] contains clean, read speech in Malay of about 140 hours. The database has already been used for conducting speech recognition research tasks (see [25], [32], [33]). A total of 199 speakers participated in the data collection where each speaker read texts from news articles for about 45 minutes. For English, we used the first release of TED-LIUM [34] corpus. The corpus contains speeches excerpted from video talks of the TED website. The transcriptions of the 120h data were generated for the International Workshop on Spoken Language Translation (IWSLT) evaluation campaign in 2011.

4.4. Cross-lingual SGMM

To obtain cross-lingual SGMM, first we trained monolingual UBMs on Malay and English data. We used training corpus available in the MASS and TED-LIUM corpora to build the models and the respective datasets contain 120h Malay and 118h English speeches. We set 600 Gaussians in the UBM and phonetic subspace dimension to 40. Then, each UBM was used in SGMM training for Iban. We applied the same training conditions used in the monolingual experiment for obtaining cross-lingual SGMM. Apart from using monolingual (English or Malay) UBM, we employed multilingual UBM in the cross-lingual method. We proposed four multilingual UBMs built using data from Iban, Malay and English. Each UBM was trained on multilingual data that contain two or three language data. The performance of these different data combinations will help us to observe which language has better impact on Iban ASR. The combinations of data are stated in Table 2b along with the multilingual SGMM results.

After evaluating the systems on the test corpus, we found that the cross-lingual approaches improved our baseline results for the very low-resource setting only (1h training data) where we gained 7% to 11% absolute WER improvement against the SGMM baseline (37.8%). Between English and Malay, the latter gave greater impact to the Iban system. For example, employing Malay UBM yielded 9.5% absolute WER improvement while using English UBM resulted 7%. In addition, using UBM trained on multilingual data that had English speech did not give good results while Iban + Malay combination was the best.

4.5. Language-specific top layer for DNN

Based on the observations of cross-lingual SGMM for Iban, Malay acoustic features are useful for improving Iban ASR. Thus, we hypothesize that DNN trained on Malay data is also beneficial for improving acoustic models in Iban system. To test this assumption, we conducted the following experiment.

We obtained two speaker adapted DNN systems; one trained on Malay training data and the other one trained on English training data. The DNN targets were 3K context-dependent triphone states for Malay and English, which

Table 2: *Performance of Iban ASR systems in terms of WER (%) for 1h and 7h systems - 1h test set - no speaker adaptation*

(a) *Monolingual Iban ASR results - no speaker adaptation*

Training approach	Amount of training data	
	1-hour	7-hour
GMM	40.3	36.0
SGMM	37.8	18.9
DNN	26.9	18.4
# of states	661	2998

(b) *Results for cross-lingual SGMM - no speaker adaptation*

Cross-lingual SGMM	Amount of training data	
	1h	7h
Using monolingual UBM:		
a. Malay	28.3	19.4
b. English	30.8	19.2
Using multilingual UBM:		
a. Iban + Malay	27.2	19.6
b. Iban + English	29.8	19.2
c. English + Malay	29.4	19.1
d. Iban + Malay + English	28.3	19.2
# of substates	805	10K

were obtained from HMM/GMM systems. For acquiring the GMM systems, we employed 39 MFCC with deltas and deltas deltas, applied LDA, MLLT and fMLLR. We trained seven-layer DNNs, each hidden layer has 1024 units. Subsequently, we removed the last layer of each DNN.

Table 3: *WERs of cross-lingual DNNs - with speaker adaptation*

DNN with lang. specific top layer	Amount of train data	
	1h	7h
a. Hidden layers from English	19.1	15.2
b. Hidden layers from Malay	18.9	15.2

Following this, we obtained DNN targets for Iban by acquiring a speaker adapted HMM/GMM system. We trained new Iban triphone models on new feature vectors by using the same feature transformation methods described above. During the LDA+MLLT training, one important trick is to use feature transforms acquired from the source (Malay or English) corpus (with large number of speakers). This is because merging DNNs with different feature transforms is not a good approach (for which we have observed no improvement). Finally, we built language-specific DNN for Iban by fine-tuning the hidden layers from Malay and English on 1h and 7h Iban training data. Then, the DNN systems were evaluated on the Iban test set. The results of the DNN systems are presented in Table 3. Applying both speaker adaptation (fMLLR) and language-specific top layer technique significantly improve our DNN baselines (reported in Table 2a, second last line). For comparison, training a monolingual and speaker adapted Iban DNN lead to 15.8% WER with 7h train condition. The results also showed few language effect (English or Malay) even if for 1h training condition, the hidden layers from English were less effective for Iban ASR.

5. Towards zero-shot ASR using a closely-related language

Our final contribution for this paper is developing a zero-shot ASR for a target language using data from a closely-related language. The zero-shot system is built through unsupervised training on Iban data. Here, we assume that we only have a language model and lexicon that are dependent to the target language. The Iban training transcripts, however, are automatically generated using a Malay acoustic model. To perform this task we built a Malay acoustic model on Malay 120h training data, using SGMM approach. Then, we employed the Malay acoustic models and the Iban LM and lexicon for decoding the Iban training corpus. The hypothesized transcripts were then directly used for training Iban ASR systems based on GMM, SGMM and DNN. All three systems were evaluated on Iban test data.

Table 4 presents performance comparison of supervised and zero-shot (unsupervised) ASR systems. Note that we did not perform cross-lingual acoustic modelling in this last experiment. In the first row of results, we present the ASR results as indicated in Table 2a. The next line shows the performance of supervised systems with speaker adaptive training applied. The SGMM and DNN acoustic models of the supervised systems were built on the GMM system which yielded the best performance in the pronunciation dictionary evaluation (last line and second column of Table 1). In general, the performance of the unsupervised system (last line) was quite close to the performance of the supervised system. We observed only 2% to 3% WER difference between each supervised and unsupervised system. The small difference suggests that the zero-shot system and the language dependent system were able to produce almost the same transcripts. However, since no difference between the performance of SGMM and DNN is observed for zero-shot ASR, we hypothesize that DNN training might be less robust to the use of noisy transcripts.

Table 4: *Performance of Iban ASR with Supervised and Unsupervised transcripts - Training on 7h of Iban speech*

ASR system (7h)	GMM	SGMM	DNN
Supervised (no spkr adapt.)	36.0	18.9	18.4
Supervised (with spkr adapt.)	19.7	16.6	15.8
Unsupervised (with spkr adapt.)	21.4	18.6	18.9

6. Conclusions

The paper demonstrates our efforts in obtaining and improving ASR for Iban, a language spoken in Malaysia. The Iban ASR corpus, which we have developed, and Kaldi scripts are available on github[4] for the speech community to use. In our work, we have applied three strategies that employed out-of-language data to build our systems: (1) semi-supervised lexicon design, (2) cross-lingual acoustic model training based on SGMM and DNN, and (3) unsupervised training of a zero-shot ASR. After using the first method, we found that it is better to start from Malay (closely-related language) than from English for building the Iban lexicon. The second approach helped us to improve our monolingual systems. For cross-lingual SGMM, we gained more WER improvements when Malay was employed and the effect was only pronounced for systems with very low-resource setting (1h). Fine-tuning hidden layers from Malay DNN also improved our DNN baselines for Iban, particularly for 1h training condition. Last but not least, the third (unsupervised) approach provided a system with promising performance. However, it seems that using automatic transcripts deserves DNN training since we observe no difference in the results of the unsupervised SGMM and DNN systems. As a future work, the use of confidence measures could help to select the best hypotheses for DNN training.

[4]https://github.com/sarahjuan/iban

7. References

[1] L. Besacier, E. Barnard, A. Karpov, and T. Schultz, "Automatic speech recognition for under-resourced languages : A survey," *Speech Communication Journal*, vol. 56, pp. 85–100, January 2014.

[2] T. Schultz and A. Waibel, "Fast bootstrapping of lvcsr systems with multilingual phoneme sets," in *Proceedings of Eurospeech*. Citeseer, 1997, pp. 371–374.

[3] ——, "Multilingual and crosslingual speech recognition," in *Proceedings of DARPA workshop on Broadcast News Transcription and Understanding*, 1998, pp. 259–262.

[4] H. Lin, L. Deng, D. Yu, Y. fan Gong, A. Acero, and C.-H. Lee, "A study on multilingual acoustic modeling for large vocabulary asr," in *Proceedings of ICASSP*, Taipei, 2009, pp. 4333–4336.

[5] Z. Wang, T. Schultz, and A. Waibel, "Towards universal speech recognition," in *Proceedings of International Conference on Multimodal Interfaces*, 2002.

[6] D. Povey, L. Burget, M. Agarwal, P. Akyazi, K. Feng, A. Ghoshal, O. Glembek, N. Goel, M. Karafiat, A. Rastrow, R. C. Rose, P. Schwarz, and S. Thomas, "Subspace gaussian mixture models for speech recognition," in *Proceedings of ICASSP*, 2010.

[7] D. Povey, L. Burget, M. Agarwal, P. Akyazi, F. Kai, A. Ghoshal, O. Glembek, N. G. M. Karafiàt, A. Rastrow, R. C. Rose, P. Schwartz, and S. Thomas, "The subspace gaussian mixture model - a structured model for speech recognition," *Computer Speech and Language*, vol. 25, pp. 404–439, 2011.

[8] L. Lu, A. Ghoshal, and S. Renals, "Cross-lingual subspace gaussian mixture models for low-resource speech recognition," in *IEEE/ACM Transactions on Audio, Speech and Language Processing*, vol. 22, January 2014, pp. 17–27.

[9] D. Imseng, P. Motlicek, H. Bourlard, and P. N. Garner, "Using out-of-language data to improve under-resourced speech recognizer," *Speech Communication*, vol. 56, no. 0, pp. 142–151, 2014.

[10] G. Hinton, L. Deng, D. Yu, A.-R. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T. S. G. Dahl, and B. Kingsbury, "Deep neural networks for acoustic modeling in speech recognition," *IEEE Signal Processing Magazine*, vol. 29, no. 6, pp. 82–97, 2012.

[11] Y. Miao and F. Metze, "Improving low-resource cd-dnn-hmm using dropout and multilingual dnn training." in *Proceedings of INTERSPEECH*, 2013, pp. 2237–2241.

[12] N. T. Vu, D. Imseng, D. Povey, P. Motlíček, T. Schultz, and H. Bourlard, "Multilingual deep neural network based acoustic modeling for rapid language adaptation," in *Proceedings of ICASSP*, 2014.

[13] J.-T. Huang, J. Li, D. Yu, L. Deng, and Y. Gong, "Cross-language knowledge transfer using multilingual deep neural network with shared hidden layers," in *Proceedings of ICASSP*, 2013.

[14] P. Swietojanski, A. Ghoshal, and S. Renals, "Unsupervised cross-lingual knowledge transfer in dnn-based lvscr," in *Proceedings of ICASSP*, 2013.

[15] G. Heigold, V. Vanhoucke, A. Senior, P. Nguyen, M. Ranzato, M. Devin, and J. Dean, "Multilingual acoustic models using distributed deep neural networks," in *Proceedings of ICASSP*, 2013.

[16] M. P. Lewis, G. F. Simons, and C. D. Fennig, *Ethnologue : Languages of the world, Seventh Edition*. SIL International, 2014. [Online]. Available: http://www.ethnologue.com

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

[18] S. R. Maskey, A. W. Black, and L. M. Tomokiyo, "Bootstrapping phonetic lexicons for language," in *Proceedings of INTERSPEECH*, 2004, pp. 69–72.

[19] M. Davel and E. Barnard, "Bootstrapping in language resource generation," in *Proceedings of 14th Annual Symposium of the Pattern Recognition Association of South Africa*, 2003.

[20] ——, "The efficient generation of pronunciation dictionaries: Human factors during bootstrapping," in *Proceedings of INTERSPEECH*, 2004.

[21] S. S. Juan, L. Besacier, and S. Rossato, "Semi-supervised g2p bootstrapping and its application to asr for a very under-resourced language: Iban," in *Proceedings of Workshop for Spoken Language Technology for Under-resourced (SLTU)*, May 2014.

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

[23] J. R. Novak. (2012) Phonetisaurus: A wfst-driven phoneticizer. available at : https://code.google.com/p/phonetisaurus.

[24] J. R. Novak, N. Minematsu, and K. Hirose, "Evaluations of an open source wfst-based phoneticezer," The Institute of Electronics, Information and Communication Engineers, PDF, General Talk No. 452, 2011.

[25] T.-P. Tan and B. Rainavo-Malançon, "Malay grapheme to phoneme tool for automatic speech recognition," in *Proceedings of Workshop of Malaysia and Indonesia Language Engineering (MALINDO) 2009*, 2009.

[26] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlíček, P. Schwarz, J. Silovskỳ, G. Stemmer, and K. Veselỳ, "The kaldi speech recognition toolkit," in *Proceedings of Workshop on Automatic Speech Recognition and Understanding*, I. S. P. Society, Ed., vol. IEEE Catalog No. : CFP11SRW-USB, December 2011.

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

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

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

[30] G. E. Hinton, "A practical guide to training restricted boltzmann machines," Dept. Computer Science, University of Toronto, UTML TR 2010-003, 2010.

[31] B. Kingsbury, "Lattice-based optimization of sequence classification criteria for neural network acoustic modeling," in *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing*, April 2009, pp. 3761–3764.

[32] X. Xiao, E. S. Chng, T.-P. Tan, and H. Li, "Development of a malay lvscr system," in *Proceedings of Oriental COCOSDA*, Kathmandu, Nepal, 2010.

[33] S. S. Juan, L. Besacier, and T.-P. Tan, "Analysis of malay speech recognition for different speaker origins," in *Proceedings of International Conference on Asian Language Processing (IALP)*. IEEE, 2012, pp. 229–232.

[34] A. Rousseau, P. Deléglise, and Y. Estève, "Ted-lium: An automatic speech recognition dedicated corpus," in *Proceedings of LREC*. European Language Resources Association (ELRA), 2012, pp. 125–129.