USING CLOSELY-RELATED LANGUAGE TO BUILD AN ASR FOR A VERY UNDER-RESOURCED LANGUAGE: IBAN

Sarah Samson Juan[1]*, Laurent Besacier*[1]*, Benjamin Lecouteux*[1]*, Tien-Ping Tan*[2]

[1]Grenoble Informatics Laboratory (LIG), Grenoble-Alpes University, Grenoble, France
[2]School of Computer Science, Universiti Sains Malaysia, Penang, Malaysia
[1]{sarah.samson-juan, laurent.besacier, benjamin.lecouteux}@imag.fr, [2]tienping@cs.usm.my

ABSTRACT

This paper describes our work on automatic speech recognition system (ASR) for an under-resourced language, namely the Iban language, which is spoken in Sarawak, a Malaysian Borneo state. To begin this study, we collected 8 hours of speech data due to no resources yet for ASR concerning this language. Following the lack of resources, we employed bootstrapping techniques on a closely-related language to build the Iban system. For this case, we utilized Malay data to bootstrap the grapheme-to-phoneme system (G2P) for the target language. We also developed several G2Ps to acquire Iban pronunciation dictionaries, which were later evaluated on the Iban ASR for obtaining the best version. Subsequently, we conducted experiments on cross-lingual ASR by using subspace Gaussian Mixture Models (SGMM) where the shared parameters obtained in either monolingual or multilingual fashion. From our observations, using out-of-language data as source language provided lower WER when Iban data is very imited.

Index Terms— automatic speech recognition, acoustic modelling, subspace Gaussian mixture model, bootstrapping grapheme-to-phoneme

1. INTRODUCTION

Speech applications have assisted the human-computer interaction for many tasks, e.g. voice command, speaker identification and speech translation systems. Today, these applications are within reach and can be found on desktop computers or mobile devices. Besides that, systems can work on multiple languages especially on languages with high amount of available data, rich in linguistic knowledge, etc. However, there are still many languages that are not yet available in these systems. Knowledge-poor and resource-scarce languages for instance, are still far behind in exploration in the speech recognition domain. The time and effort to build systems for new languages is costly with several constraints to tackle such as no pronunciation dictionary, lack of speaker diversity in the collected speech and unstable orthography system [1].

Nevertheless, we believe that it is possible to use similar linguistic knowledge that exist between languages as a starting point to develop data for ASR, for example, the pronunciation dictionary or acoustic models. Bootstrapping a G2P [2] is a strategy to reduce effort of producing phonetic transcriptions for all of the words in a vocabulary from scratch. Commonly, this semi-supervised method requires a transcript that contains words and the respective pronunciations in a target language, usually created by a native speaker or a linguist. This transcript then becomes the seed for a pronunciation model. The model is then used to predict new entries in the vocabulary and post-editing can be carried out later, if needed. The process of updating the model can be repeated by adding the post-edited list into the model. This strategy saves time and effort to build pronunciations for a large vocabulary word list. We have shown in our previous work ([3], [4]) that it is also feasible to prepare a pronunciation model for a target language from an existing one in a similar (source) language. We experimented on using a grapheme-to-phoneme system (G2P) of Malay, a language from the Austronesian language family, to produce a base pronunciation transcript for Iban, a language from the same family. We post-edited the outputs and used the improved version later as a seed lexicon for the Iban G2P. The first contribution of the present paper consists in evaluating the impact of the source G2P (e.g. similar language like Malay, different language such as English, grapheme-based approach with no knowledge at all) on ASR accuracy.

Concerning feature extraction or acoustic modelling, studies have demonstrated that cross-lingual acoustic approaches can help to boost the accuracy of low-resource ASR (see [5], [6], [7]). Adapting the acoustic models that are trained from out-of-language data to a system that has limited amount of training data proves to be an effective approach to improve monolingual system results. However, the multilingual acoustic modelling approaches described above require a mapping between (multilingual) source phone units and their target language counterpart. This stage might be tricky, especially for very under-resourced languages that are poorly described. This is why recent studies on cross-lingual

acoustic modelling based on subspace Gaussian mixture model (SGMM) seem very promising for speech recognition in limited training data conditions ([8], [9]). With SGMMs, units distributions are all derived from a common GMM called UBM (Universal Background Model). This UBM can be trained on a large amount of un-transcribed data and recent cross-lingual approaches attempted to train SGMMs using cross-lingual or multi-lingual approaches (UBM trained on one or several languages different to the target language). Unlike the cross-lingual technique proposed by Schultz et al. ([6], [7], [10]), the globally shared parameters in SGMM approach do not need knowledge about the phone set used in source language(s). Thus, SGMMs were very recently used to train a multilingual subspace, as shown in the work of Lu et al. [8]. In addition, the use of a UBM trained on many different speakers can also help to handle the lack of speaker diversity found in transcribed speech resources for very low-resourced languages (where only few speakers are generally recorded).

This paper focuses on ASR for Iban, a very under-resourced language. Recently, we used a Malay G2P to help build an Iban G2P for ASR in view of several similarities between the two languages (similar orthography system, pronunciations). In this paper, we present our additional experiments on G2P by evaluating Iban ASR with pronunciation dictionaries created by out-of-language G2Ps (English and Malay) as well as a knowledge-free G2P (grapheme-based). Apart from that, we investigate cross-lingual effects to Iban ASR when training data are limited (with two different training data size; 1 hour and 7 hours).As the acoustic properties of a source language data can be directly applied in SGMM training for any target language data, we use this opportunity to employ data from two languages, a similar and reasonably well-resourced one: Malay - and a different but very well-resourced: English.

The remainder of this paper explains further details about Iban resources and the techniques that we applied to build the Iban ASR. Section 2 describes the target language briefly and reports available data for ASR experiments while Section 3 presents the bootstrapping of G2P for pronunciation modelling. Section 4 presents our experiments using out-of-language data while Section 5 displays the results. Finally, Section 6 concludes the paper and provides perspectives.

2. THE IBAN LANGUAGE AND RESOURCES

2.1. Iban in brief

Iban is a regional language mainly spoken in Sarawak, Malaysia, mostly by the Iban community. The Ibans consists of 30.3% of the total population in the state [11]. The language system is similar to Malay in terms of phonology, morphology and syntax. Both languages belong to the Malayo-Sumbawan branch (Austronesian language family)[12] and they are written using Latin alphabets. It is known that Malay and Iban share words and there are many Malay words integrated (borrowed) into Iban for vocabulary growth[13]. With this connection between Iban and Malay, we try to take advantage of Malay, a reasonably well-resourced language to assist in the creation of Iban inputs for an ASR system. (To see examples of Malay and Iban words and pronunciations, refer to [3]).

2.2. Speech and transcript

We collected news data from a local radio station. Almost eight hours of news data was provided by *Radio Televisyen Malaysia* (RTM). The data was later transcribed by eight Iban speakers using Transcriber ([14]). The signals were segmented according to sentences and noise (page turns, music, etc) was discarded. After this process, we have more than 3K sentences uttered by 25 speakers. From here, we split the data into two sets; train and test. Table 1 shows further details on the speech corpus.

Table 1. Amount of Iban transcribed speech (training and testing)

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

2.3. Text for language modelling

We found an online news website[1] that publishes Iban articles over the past few years. From this website, texts dated from 2009 to 2012 were extracted through web crawling approach. In total, we have 7K articles on sports, entertainment and general matters. Subsequently, we conducted text normalization on the data using the following procedure : (1) remove HTML tags, (2) convert dates and numbers to words (e.g: 1973 to *sembilan belas tujuh puluh tiga*), (3) convert abbreviations to full terms (e.g: Dr. to *Doktor*, Prof. to *Profesor*, Kpt. to *Kapten*), (4) split paragraphs to sentences, (5) change uppercase characters to lowercase and (6) remove punctuation marks (except hyphen / '-'). After completing these steps, there are 2.08M words and 37K unique words identified. Using SRILM toolkit [15], we developed a trigram language model with modified Kneser-Ney discounting applied. The perplexity of the model was then measured based on the speech transcript. We achieved a perplexity of 158 (2.3% OOV rate) for the Iban language model correspondingly.

3. SEVERAL STRATEGIES FOR OBTAINING IBAN PRONUNCIATION DICTIONARY

We obtained a Malay pronunciation dictionary from the MASS corpus [16]. The dictionary was used for a Malay

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

Table 2. Iban ASRs performances (WER%) using different pronunciation dictionaries (7hr training data)

Training approach	Dictionary				
	Grapheme	English G2P	Malay G2P	Iban G2P	Hybrid G2P
Monophone	40.04	48.8	42.17	41.79	41.97
Triphone + Δ + Δ	33.85	39.91	36.47	36.98	36.77
+ MLLT + LDA	26.52	30.20	27.24	27.71	26.80
+ SAT(fMLLR)	21.43	28.96	20.82	21.90	**20.60**

ASR where a total of 76K Malay pronunciations are available. Then, we trained a **Malay G2P** on Phonetisaurus[2][17], an open source G2P tool based on Weighted Finite States Transducers (WFSTs). For training the Malay G2P, we chose a subset of 68K pronunciations. The G2P was then used to phonetize 1K Iban words for obtaining a base pronunciation transcript. Following that, an Iban native speaker (first author of this paper) corrected the outputs of the system and we took the post-edited pronunciations to build an **Iban G2P**. After that, both systems were applied to another 1K words from the Iban word list and the outputs were post-edited. Later, we evaluated both generated and reference transcripts and found that the Malay G2P can phonetize Malay-Iban (same surface forms) more accurately than the Iban system, while, the Iban system works better for pure Iban (not-shared with Malay). Afterward, we phonetized the whole Iban lexicon based on the following approach (later called **Hybrid G2P**): the Malay G2P phonetizes all Malay-Iban while the Iban G2P phonetizes all pure Iban words. Consequently, our 37K word lists was phonetized using these 3 G2P systems (Malay, Iban, Hybrid). The best performing system (Hybrid G2P) obtained 8.1% PER and 29.4% WER from a 2K random outputs assessment. More details on the investigation of Malay and Iban pronunciations can be found in [3] and [4].

Apart from these three G2P systems (Malay, Iban and Hybrid), we took the chance to explore 2 other phonetizers, a **grapheme-based** one (using no knowledge) and an **English G2P**. The grapheme-based phonetizer was built based on Malay segmentation rules [18] while the English G2P is the demo system built from English CMU[3] pronunciation list for Phonetisaurus.

4. ASR EXPERIMENTS USING OUT-OF-LANGUAGE DATA

We conducted the experiments on Kaldi [19], a speech recognition toolkit based on FSTs. We focused on two types of assessments. First, we aimed to test all five dictionaries separately on Iban ASR. After obtaining baseline results, our second investigation involved testing SGMM for Iban. We set two levels of data sparseness on Iban data; one with 7 hour training data and the other with 1 hour training data (subset of the 7 hour). For this SGMM evaluation, we employed pronunciation dictionary that performs best in the first experiment (based on Hybrid G2P). All Iban systems used the trigram Iban language model that we acquired before.

[2]available at https://code.google.com/p/phonetisaurus

[3]available at https://cmusphinx.svn.sourceforge.net/svnroot/cmusphinx

4.1. Impact of the Pronunciation Dictionaries

We used 13 MFCCs and Gaussian mixture models (GMM) for monophone and triphone trainings on 7 hours Iban data. For triphone training, we applied 2,998 context-dependent states and 40K Gaussians. We also implemented delta delta coefficients on the MFCCs, linear discriminant analysis (LDA) transformation and maximum likelihood transform (MLLT) [20], and, speaker adaptation based on feature-space maximum likelihood linear regression (fMLLR) [21]. We applied each dictionary separately for training the acoustic model (AM) resulting five Iban recognizers for evaluation (Grapheme-based, English G2P, Malay G2P, Iban G2P and Hybrid G2P). Section 5 will explain about the ASR results.

4.2. Using SGMM Acoustic Modelling

The GMM and SGMM acoustic models are similar where each emission probability of each HMM state is modelled with a Gaussian mixture model. In the SGMM approach, instead of estimating GMM parameters directly from the training data like in the conventional approach, the Gaussian means and mixture component weights are generated from the phonetic and speaker subspaces along with a set of weight projections. The SGMM model is described in the following equations [22]:

$$p(\mathbf{x}|j) = \sum_{m=1}^{M_j} c_{jm} \sum_{i=1}^{I} w_{jmi} \mathcal{N}(\mathbf{x}; \mu_{jmi}, \mathbf{\Sigma}_i) \quad (1)$$

$$\mu_{jmi} = \mathbf{M}_i \mathbf{v}_{jm}, \quad (2)$$

$$w_{jmi} = \frac{\exp \mathbf{w}_i^T \mathbf{v}_{jm}}{\sum_{i'=1}^{I} \exp \mathbf{w}_{i'}^T \mathbf{v}_{jm}} \quad (3)$$

where $\mathbf{x} \in \mathbb{R}^D$ denotes the D-dimensional feature vector, $j \in \{1..J\}$ is the HMM state, i is the Gaussian index, m is the substate and c_{jm} is the substate weight. Each state j is associated to a vector $\mathbf{v}_{jm} \in \mathbb{R}^S$ (S is the phonetic subspace dimension) which derives the means, μ_{jmi}

and mixture weights, w_{jmi} and it has a shared number of Gaussians, I. The phonetic subspace $\mathbf{M}_i$, weight projections $\mathbf{w}_i^T$ and covariance matrices $\mathbf{\Sigma}_i$, i.e., the globally shared parameters $\mathbf{\Phi}_i = \{\mathbf{M}_i, \mathbf{w}_i^T, \mathbf{\Sigma}_i\}$ are common across all states. These parameters can be shared and estimated over multiple language data.

To implement the SGMM training, we used the same decision trees as the ones being used in the monolingual system. A generic mixture of I Gaussians, denoted as Universal Background Model (UBM), models all the speech training data for the initialization of the SGMM. It is important to note that we did not apply speaker adaptive training in the SGMM experiments. During training, we used different number of substates for both monolingual and crosslingual SGMM to study its impact on the SGMM modelling performance.

4.2.1. Monolingual SGMM

The 7-hour GMM system that gave the best result in the pronunciation dictionary evaluation (using Hybrid G2P) was chosen for SGMM training. Another Iban ASR system was built using only 1 hour (1h) data to show limited training data. For the 1h system, we chose speeches uttered by four female and three male speakers and used the Hybrid G2P pronunciation dictionary. We obtained the context dependent model for the 1h-system using 661 states and 5K Gaussians. Then, we trained UBM from the 7h and 1h systems by setting I and S to 600 and 40 respectively. Subsequently, SGMM training was done using the same decision tree obtained in earlier GMM training step (recall that this tree is different for each training condition: 1h = 661 states/5K gaussians and 7h = 2,988 states/40K gaussians).

4.2.2. Cross-lingual and Multilingual SGMMs

This ASR experiment involved obtaining SGMM shared parameters in cross-lingual (using out of language data to train the UBM) and multilingual (using 2 or more languages to train the UBM) fashion. To prepare this investigation, we used Malay and English data from the MASS corpus (read speech) [16] and TED corpus [23]. UBM models, but also full ASR systems were trained using 120 hours (175 speakers) of Malay and 118 hours (666 speakers) of English. We also went through the same training process as the one described for Iban ASR and observe the systems' performances on 20-hour Malay and 4-hour English test data. This was a way for us to assess how the out-of-language data affects our SGMM experiments.

Finally, we developed two cross-lingual (from Eng UBM referred to as ENG_cl ; or Malay UBM referred to as MY_cl) and four multilingual systems for SGMM training. Our multilingual data compositions (pool existing training data) were as follows :(a) Eng + Malay UBM (referred to as EM_mul),(b) Eng + Iban UBM (referred to as EI_mul),(c) Iban + Malay UBM (referred to as IM_mul) and (d) Eng + Malay + Iban UBM (referred to as EIM_mul). Once all the UBMs were obtained (either in a cross-lingual or multilingual fashion), the other steps of the SGMM training took place and they were the same as for the monolingual SGMM design (SGMM subspace parameters estimated on the available training data 1h or 7h). The number of UBM Gaussians (600) and phonetic subspace dimension (40) followed the previous setting.

5. EXPERIMENTAL RESULTS

We report the ASR performance results based on the two experiments described in the preceding section. Several language model weights were applied for each recognition experiment and we systematically picked up the best one to be reported in this paper.

5.1. Baseline GMM modelling

Table 2 summarizes our baseline results based on five different pronunciation dictionaries. On average, using monophone models provided us 43.4% WER while applying triphone models with several features can reduce the WERs to half of the monophone average result, giving 23% WER. Hybrid G2P system provides the highest accuracy among the rest (20.60% WER). However, this is only a slight improvement from systems with Malay or Iban based dictionaries. Eventhough English G2P resulted the worst among all the systems, the performance is only 8% different (28.96% WER) than the other systems. This shows that using an out-of-language G2P can be a decent starting point to develop ASR for a very under-resourced language. As expected, the ASR performance is better if the out-of-language language G2P comes from the same language group (Malay, 20.82% WER) than from a different language group (English, 28.96% WER). Moreover, using a grapheme-based system is also a very good option since it gave similar results with systems using Malay or Iban based G2P.

5.2. SGMM systems

5.2.1. Baseline monolingual experiments

Table 3. Baseline Iban ASR results (WER%) for monolingual GMM and SGMM approaches

Training approach	IB System	
	1-hour	7-hour
GMM	41.17	36.77
SGMM (no speaker transform)	38.79	20.56
# of states	661	2998
# of substates	805	4111

Table 3 presents the monolingual GMM and SGMM baseline results for Iban (IB). Note that for a common training

condition (1h or 7h) both systems used the same decision trees. Furthermore, we utilized Hybrid G2P dictionary in both systems. We can observe that the SGMM system outperformed the GMM system even when the subspace parameters were estimated on a very limited data as observed from the 1-hour condition. It managed to reduce up to 2.33% from the GMM result. For comparison, the Malay (MY) and English (ENG) ASRs baseline results are also presented in Table 4. We found that the SGMM systems also outperformed the GMM systems for the two languages. In the cross-lingual experiments presented in next section, the UBMs from Malay and English SGMMs are used.

Table 4. Baseline English and Malay ASR results (WER%) - systems were trained on the data we use to train our UBMs

Training approach	System(test size)	
	MY (20hrs)	ENG (4hrs)
GMM (Triphone + LDA + MLLT)	7.05	29.88
SGMM (with speaker transform)	4.31	22.25

5.2.2. Cross-lingual and multilingual experiments

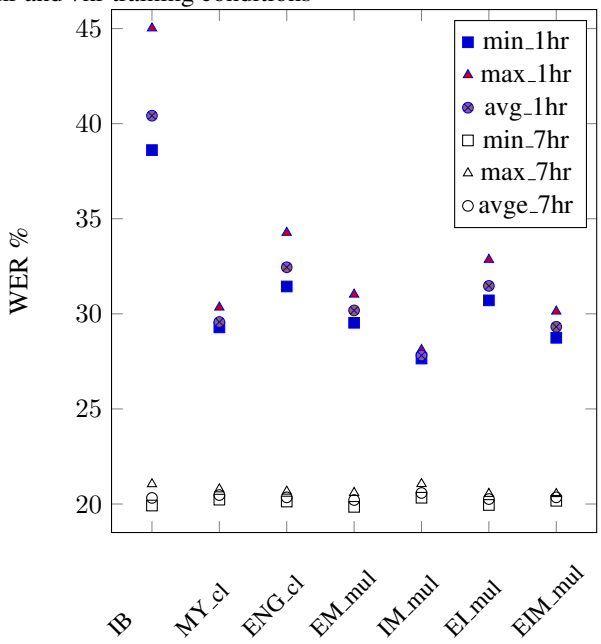
Figure 1 shows our results on monolingual, cross-lingual and multilingual SGMM systems. In the graph, we present the minimum, average and maximum WER values from our observations after applying different number of substates. For the 1h system evaluation, we used substate values ranging from 800 to 8700, while for the 7h system, we used 4200 to 56000 substates.

From this graph, we can observe that the WER results of the 1h system (blue or dark shaded plots) were greatly improved when cross-lingual SGMM (ENG_cl and MY_cl) applied. In fact, training SGMM parameters from an out-of-language UBM significantly reduced the WER from an SGMM monolingual (IB) baseline. As for the pronunciation dictionary experiments, Malay (same language group) was better than English (different language group) as an out-of-language data for cross-lingual experiments. The multilingual experiments (EM_mul, IM_mul, EI_mul, EIM_mul) are also better than monolingual SGMM but it is difficult to find the optimal language combination: further improvements are shown when pooling Iban and Malay data for UBM training, but slight degradation is observed when pooling Iban and English. Overall, for the 1h training condition, the best SGMM system managed to reduce 20% WER from the monolingual GMM system. As for the 7-hour system, the cross-lingual SGMM results did not show much improvement (nor degradation) compared to the monolingual SGMM.

6. CONCLUSIONS AND FUTURE WORK

We have demonstrated our work on building an ASR for Iban, a very under-resourced language. We showed that

Fig. 1. Min, max and average results of Iban (monolingual, cross-lingual and multilingual) SGMM experiments based on 1hr and 7hr training conditions



using data from a closely-related language can quickly build an Iban system. During the course of development, we created an Iban pronunciation dictionary via bootstrapping strategy based on Malay data. In addition, different dictionary versions were produced using several approaches which were then tested on the Iban ASR. We found that the hybrid version (Hybrid G2P) gave the lowest WERs (20.6%). Then, our study focused on improving the GMM system result using SGMM approach. We investigated cross-lingual SGMM by obtaining UBMs in monolingual/multilingual fashion that were later applied to the Iban AM training. Our results showed that using English and Malay as source language data manage to reduce WER (from monolingual SGMM) significantly for the Iban 1-hour system. We plan to further explore cross-lingual approaches that can help to improve current results.

7. REFERENCES

[1] 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.

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

[3] 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.

[4] Sarah Samson Juan, Laurent Besacier, and Solange Rossato, "Semi-supervised g2p bootstrapping and its application to asr for a very under-resourced language: Iban," in *Workshop for Spoken Language Technology for Under-resourced (SLTU)*, May 2014.

[5] Samuel Thomas, Sriram Ganapathy, and Hynek Hermansky, "Cross-lingual and multi-stream posterior features for low resource lvcsr systems," in *Proc. INTERSPEECH*, 2010, pp. 877–880.

[6] Tanja Schultz and Alex Waibel, "Multilingual and crosslingual speech recognition," in *Proc. DARPA workshop on Broadcast News Transcription and Understanding*, 1998.

[7] Tanja Schultz and Alex Waibel, "Language-independent and language-adaptive acoustic modeling for speech recognition," *Speech Communication*, vol. 35 : 1, pp. 31–52, 2001.

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

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

[10] Tanja Schultz, "Globalphone: a multilingual speech and text database developed at karlsruhe university," in *Proc. ICLSP*, 2002, pp. 345–348.

[11] "Negeri sarawak : Total population by ethnic group, sub-district and state," Tech. Rep., Malaysian Statistics Department, Malaysia, 2010.

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

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

[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] 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.

[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, 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.

[18] Tien-Ping Tan and Bali Rainavo-Malançon, "Malay grapheme to phoneme tool for automatic speech recognition," in *Proc. Workshop of Malaysia and Indonesia Language Engineering (MALINDO) 2009*, 2009.

[19] Daniel Povey, Arnab Ghoshal, Gilles Boulianne, Lukàš Burget, Ondřej Glembek, Nagendra Goel, Mirko Hannemann, Petr Motlíček, Petr Schwarz, Jan Silovskỳ, Georg Stemmer, and Karel Veselỳ, "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.

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

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

[22] Daniel Povey, Lukàš Burget, Mohit Agarwal, Pinar Akyazi, Feng Kai, Arnab Ghoshal, Ondřej Glembek, Nagendra Goel Martin Karafiàt, Ariya Rastrow, Richard C. Rose, Petr Schwartz, and Samuel Thomas, "The subspace gaussian mixture model - a structured model for speech recognition," *Computer Speech and Language*, vol. 25, pp. 404–439, 2011.

[23] Anthony Rousseau, Paul Deléglise, and Yannick Estève, "Ted-lium: An automatic speech recognition dedicated corpus," in *Proc. LREC*. European Language Resources Association (ELRA), 2012, pp. 125–129.