Jurnal Teknologi

Full Paper

# A System Combination for Malay Broadcast News Transcription

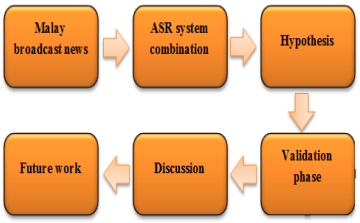
Zainab A. Khalaf[a,b*], Tien-Ping Tan[a], Li-Pei Wong[a], Basem H. A. Ahmed[c]

[a]School of Computer Sciences, Universiti Sains Malaysia (USM) 11800 Penang, Malaysia
[b]School of Computer Sciences, Basra University, Basra, Iraq
[c]School of Computer Sciences, Al Aqsa University, Gaza, Palestine

**Article history**
Received
*15 May 2015*
Received in revised form
*1 July 2015*
Accepted
*11 August 2015*

*Corresponding author
Zainab_ali2004@yahoo.com

**Graphical abstract**




**Abstract**

In this paper, we propose a post decoding system combination approach for automatic transcribing Malay broadcast news. This approach combines the hypotheses produced by parallel automatic speech recognition (ASR) systems. Each ASR system uses different language models, one which is generic domain model and another is domain specific model. The main idea is to take advantage of different ASR knowledge to improve ASR decoding result. It uses the language score and time information to produce a 1-best lattice, and then rescore the 1-best lattice to get the most likely word sequence as the final output. The proposed approach was compared with conventional combination approach, the recognizer output voting error reduction (ROVER). Our proposed approach improved the word error rate (WER) from 33.9% to 30.6% with an average relative WER improvement of 9.74%, and it is better than the conventional ROVER approach.

*Keywords:* System Combination, ROVER, Bahasa Malayu, Broadcast News



## 1.0 INTRODUCTION

Broadcast news plays a very important role in the lives of modern humans. Its main objective is to keep viewers informed about the latest developments, events and issues occurring throughout the world. The amount of news broadcasted from media devices such as radio, television, and the Internet continues to grow steadily. Thus, there is the need for systems capable of processing (e.g. indexing, summarizing, translating etc.) the digital content of broadcast news effectively and efficiently in text. Nevertheless, before other language processing tasks can be performed, the speech content in the news have to be first transcribed. The process of transcribing broadcast news is carried out by an automatic speech recognition (ASR) that renders speech into a written text [1, 2, 3].

The state of the art automatic speech recognition system employed a statistical framework. Statistical ASR aims to decode a given acoustic observation X to the corresponding word sequence W' = {$w_1$, $w_2$, . . .,$w_n$} that has the maximum expected posterior probability P(W | X),

$$W' = arg \max_{w} \; P(W|X) \quad (1)$$

$$W' = arg \max_{w} \; P(W)P(X|W) \,/P(X) \quad (2)$$

The term P(X) can be removed since the value constant, and simplify as follows.

$$W' = arg \max_{w} \; P(W)P(X|W) \quad (3)$$

where P(W) is defined by a language model and P(X | W) is defined by a pronunciation dictionary and an acoustic model. The speech decoding process can be implemented based on single-pass by solving the formula above, or it can be implemented with a multi-pass system combination. ASR output is typically a single-pass best hypothesis. However, multipass ASR system is often used to improve the accuracy of the ASR system using different knowledge sources because ASR system still produce substantial errors due factors such as data quantity and environmental conditions [4].

On the other hand, ASR system combination uses more than one ASR systems to take advantage of different techniques and knowledge in different ASR. ASR system combination is an approach that combines the decoding outputs of two or more ASRs to estimate the most likely hypothesis for a speech utterance. ASR system combination has shown to improve the word error rate (WER) of an ASR system, and it has many advantages, such as:

1. Multi-pass system can run in parallel scheme.
2. Multi-pass system combination can reduce the search space complexity.
3. Multi-pass system combination is flexible system.
4. Multi-pass system combination can use different knowledge and thus, is less expensive with respect to the training effort.

In this paper, we propose a system combination approach. The proposed method combines multiple ASRs that use different language models. A specialized language model is used in one of the ASRs that has been trained using in-domain data. Our idea is that the rescoring will produce a more accurate hypothesis by using a general language model and a domain specific language model in two different ASRs because, generic knowledge and domain specific knowledge are used in transcribing the broadcast news.

## 2.0 RELATED WORK

Nowadays, there are a lot of devices that have been embedded with speech to text capabilities. This can be seen most notably is on mobile phones and smart television. An automatic speech recognition system decodes speech to text. Decoding error however is still one of the issues when it comes to ASR system before it can be widespread applied in any devices or systems. One important area is in automatic broadcast news transcription.

There are few methods to improve the decoding of ASR system. An approach that we going to explore in this paper involves using system combination approach. In general, system combination approach involves the use of more than one ASR system to improve the decoding. ASR system combination systems can be divided depending on the place of combination, which are prior decoding, during decoding and post decoding. Prior decoding approach uses a combination of different hybrid systems. These systems make utilize of several input representations and whose outputs are joined before decoding. Example of prior decoding combination are feature-level combination and posterior-level combination [5]. This method considers a very useful and flexible approach in automatic speech recognition system. However, this method cannot be easily applied to cases that are combining information across different recognition systems.

On the other hand, during decoding combination integrates several ASR search space (*i.e.* word lattices) to enhance the speech recognition output. The frame WER (fWER) decoding combination [6, 7] and search space integrated [8] are models of this combination type. The major advantage of this approach is the direct integration of the hypothesis spaces based on time information. However, using this approach increases the search space for decoding, which can increase the processing time considerably.

Finally, the post decoding system combination approach uses rescoring module to combine possible hypotheses, which normally are n-best hypotheses or word lattice produced by ASR systems to reevaluate the hypotheses using possible additional knowledge sources. The benefit of the approach is each ASR could possibly be run in parallel in different computer. The rescoring is often very fast. One example of commonly used post decoding system used is ROVER system [9]. ROVER system reevaluates 1-best hypotheses by means of voting or confidence scores [10, 11]. The voting process is performed with a simple weighted vote that is applied to the n single-system hypotheses, and also taking into account the ASR system posteriors [5]. The evaluation scores are diverse, but they are all used in a general formula for rescoring:

$$Score(w) = \alpha\left(\frac{N(w,i)}{Ns}\right) + (1-\alpha)CS(w,i) \quad (4)$$

where *i* is the current position in the WTN, *N(w,i)* is the word *(w)* frequency at the location *i*, *Ns* is the number of combined systems, and *CS(w,i)* is the word confidence value at the position *i*. The parameter a is set to be the exchange between using the word frequency and the confidence scores [9]. Figure 1 depicts the ROVER system architecture.

Diverse systems generate different errors; thus, for the three hypotheses developed from different ASR systems, it can be assumed that:

Hyp1: "*maka stmp dalam skop komposit*"
Hyp2: "*markah skmp dalam @ komposit*"
Hyp3: "*maka skmp @ skop komposit*"

Figure 2 shows an example of ROVER result of three ASR system hypotheses (Hypothesis 1, 2 and 3). First, the word transition network (WTN) is created from these hypotheses which associated with confidence scores that computed using CMUtools [12]. The ROVER system chooses the candidate word with the highest confidence score as the final hypothesis output. The final hypothesis obtained from ROVER system is "*maka skpm dalam skop komposit*." While, the correct human translation of the speech (manual reference) is "*markah skpm dalam skor komposit*."

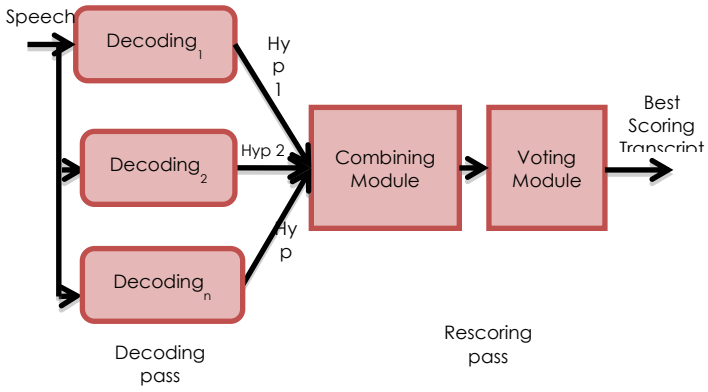



**Figure 1** The ROVER system architecture [9]

| | | | | | |
|---|---|---|---|---|---|
| **Hypothesis 1** | maka | stpm | dalam | skop | komposit |
| **Hypothesis 2** | markah | skpm | dalam | @ | komposit |
| **Hypothesis 3** | maka | skpm | @ | skop | komposit |
| **Word transition network (WTN)** | maka, 0.77 stpm, 0.1 dalam, 0.67 skop, 0.87 komposit, 1 markah, 0.23 skpm, 0.9 @, 0.33 @, 0.13 | | | | |
| **ROVER result** | maka | skpm | dalam | skop | komposit |

**Figure 2** An example of ROVER result of three ASR system hypotheses

ROVER is generally restricted to 1-best hypotheses; but it is possible to use n-best list from each ASR system [13]. Various improvements have been proposed to ROVER that are evaluated based on machine learning algorithms [14, 15], knowledge dependent language model with voting [10], optimal weighting of ASR system hypotheses based on confidence measures and others. The consistency and flexibility of post decoding system combination is due to its capacity to properly exploit multiple parallel ASR system. However, the major difficulty associated with this type stems from the fact the ASR hypothesis are integrated at word-level.

This paper proposes a post decoding system combination approach. This combination merges two ASR hypotheses generated from different ASR systems that utilize two different language models. The idea is to employ a domain specific language model on an ASR and a generic language model in two ASR systems.

This paper is organized as follow. Section 3.0 describes the proposed system. Section 4.0 introduces the speech corpora. Text corpora will explain in section 5.0. Experimental results will discuss in section 6.0. Finally, section 7.0 presents our conclusions.

## 3.0 PROPOSED POST DECODING SYSTEM COMBINATION

A good language model (LM) becomes very important as it must predict how words may be joined together to form a sentence [16]. We propose a post decoding system combination approach that takes advantage of the strength of generic and specific domain language model from two ASR systems. The combination is done using 1-best hypothesis from both ASRs, where the time and scores information in each of the 1-best hypothesis are combined to create a graph. The graph is then re-evaluated to find the most probable hypothesis. Figure 3 illustrates the post decoding system.

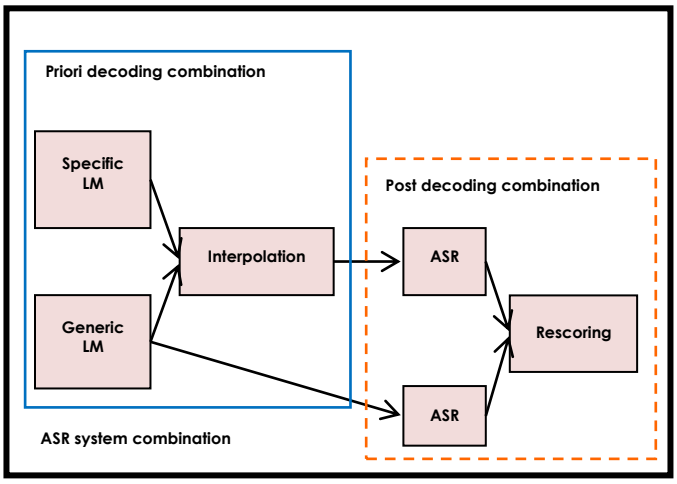



**Figure 3** The post decoding system

In the following subsections, we will explain the post decoding components in more details.

### 3.1 Priori-decoding Combination

A language model needs to train on a large corpus to estimate robust n-gram values. Ideally, the training text corpus should match the test, because n-gram is domain dependent. Therefore, a domain specific (in-domain) data is useful to adapt generic language model (out-of-domain). According to the difference between the topics, the frequency of specific n-grams can also greatly differ based on topics. For example, a commentary from a football match might have a significantly higher occurrence of the n-gram "*penalti menembak keluar*" compared with other topics such as politics. However, often domain specific knowledge is often limited compared to generic knowledge. To address this problem, we used interpolation to interpolate the generic and specific domain language models. Figure 4 shows the language model interpolation approach.

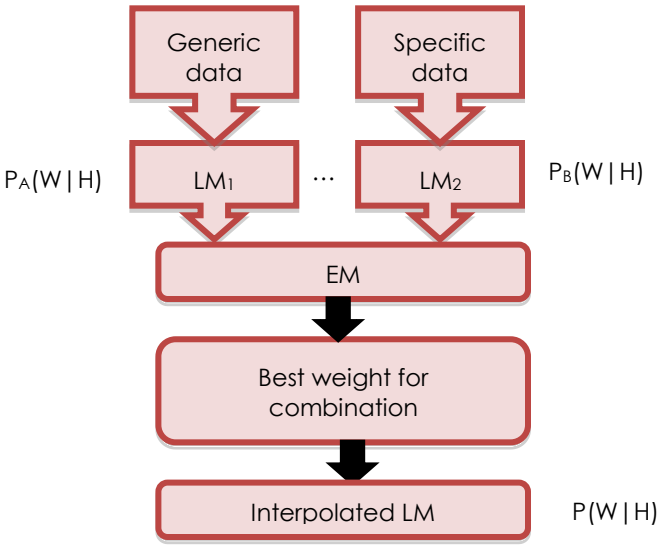



**Figure 4** General framework for language model interpolation approach

The n-gram language models $P_B(W|H)$ (computed and created from a small adaptation in-domain corpus) and $P_A(W|H)$ (computed and created from an out-of-domain corpus) are combined by simply adding together the weighted likelihoods of the two different models. Equation (5) presents the interpolation formula:

$$P(W|H) = (1-\lambda)\, P_A(W|H)\, P_B(W|H) \quad (5)$$

$\lambda$ denotes the interpolation coefficient that takes any value between 0 and 1. The expectation-maximization (EM) technique can be utilized to compute optimal weights $\lambda$ by maximization of the likelihood of the data [17]. The optimal weights $\lambda$ for a interpolation combination of the two models are calculated using the interpolation tool in the CMUtoolkit.

In the current study, two language models were used. The first one was created from large generic text corpus, which is explicitly used for training general purpose language model. The second language model was created by interpolating the large generic text corpora and a small specific corpus using an interpolation approach.

### 3.2 Post Decoding Combination

The post-decoding system combination uses two parallel ASR systems during decoding. Each ASR system produces a 1-best lattice hypothesis which will be combined and re-evaluated based on time information.

#### 1) Decoding Pass

In the decoding pass, two parallel automatic speech recognizers for Malay speech decode a speech utterance; in this case, they decode the spoken broadcast news to text using the sphinx 3 ASR system with the maximum a posteriori (MAP) and maximum likelihood linear regression (MLLR) algorithms to improve the ASR acoustic model. Each recognizer produces a most probable word sequence using the same pronunciation and acoustic models but a different language model. The Malay 1 Automatic Speech Recognizer ($M_1$ ASR) and the Malay 2 Automatic Speech Recognizer ($M_2$ ASR) each produce a hypothesis, given a speech utterance. The ASR produces a 1-best lattice as an output. In the 1-best lattice, the hypothesis contains the most probable word sequence of the speech utterance with detailed information about the acoustic score, language score, starting frame, and ending frame for words.

#### 2) Rescoring Pass

In this pass, the output from the parallel speech recognition, i.e., the 1-best lattice hypothesis is rescored in the rescoring module. Our proposed rescoring joins the hypotheses based on the information about the start and end frames to become a single word lattice.

Figure 5 shows an example of 1-best lattice rescoring. The detail of the rescoring module is as follows. The boxes present the words that appear in the 1-best lattices. The words in the lattices are joined based on the frame boundaries (BW), where a word "$W_n$" in a lattice $L_i$ is connected to another word "$W_m$" in another lattice $L_j$ if the start frame of the word "$W_m$" is between the first end frame (FE) and the last end frame (LE) of the word "$W_n$". Thus, if the first end frame of the word "$W_n$" is $W_{FE}$, the last end frame of the word "$W_n$" is $W_{LE}$, and the start frame of the word "$W_m$" is $W_S$, then the word "$W_n$" is connected to the word "$W_m$" if $W_{FE} \leq W_S \leq W_{LE}$. (3.1)

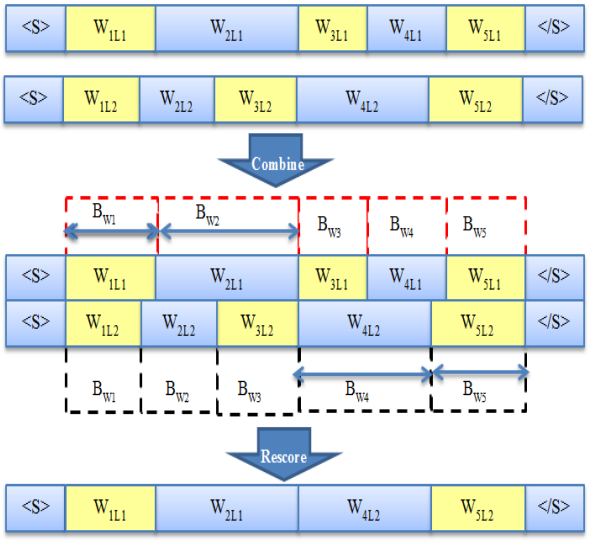



**Figure 5** 1-best lattice rescoring steps

Language model scores must be recalculated for new connecting edges that are formed as a result of connecting two words. The trigram language model score for the new connecting edge comprises the interpolation of the n-gram language model used for both words. If any trigram for the word sequence is not found, a bigram or unigram is applied using the back-off model. If there is a priori knowledge of the language of the utterances, an interpolation weight can be set. Finally, the path with the highest total acoustic and language model score can be calculated using the Viterbi algorithm.

We use Min-Max normalization [18] to rescale the weights of one language model from its range of values to the second language model range of values. Figure 6 shows the normalization of the $M_2$ scores to the $M_1$ scores, in which all $M_2$ scores are rescaled to $M_1$ scores.

To normalize $LM_1$ scores to $LM_2$ scores, first, the smallest and largest scores in both language models are determined. Let $LM_{1S}$ and $LM_{1L}$ be the smallest and largest scores, respectively of $LM_1$, while $LM_{2S}$ and $LM_{2L}$ are the smallest score and largest score, respectively, of $LM_2$. All scores $LM_{1w}$ of $LM_1$ used in the rescoring are normalized using equation (6) [18].

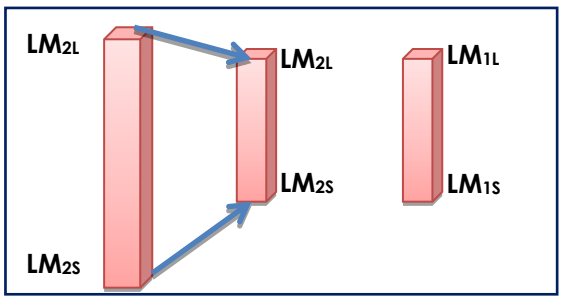



**Figure 6** Language model score normalization [16]

Figure 7 shows a two-pass rescoring system that combines two 1-best lattices ($L_1$ and $L_2$) from two ASR systems. The first ASR system uses the language model that is created from original data (generic data), while, the second ASR system uses the interpolated language model that is created from both in-domain data and generic data. The small orange rectangle box refers to the word start frame, while the dark blue text box represents the word end frame. The dotted lines represent the new connecting edges. The original edges for $L_1$ are: {undi→di, di→meri, meri→lima, lima→melaka}, and for $L_2$, they are {undi→de, de→merlimau, merlimau→melakar}. Therefore, the new edges that are needed for rescoring are {di→merlimau, de→lima, lima→melakar, merlimau→melaka}. As stated before, the language model score for the new connecting edges needs to be computed according to previous details. The final hypothesis obtained is "undi di merlimau melaka" as shown in Figure 7.

## 4.0 SPEECH CORPORA

We carried out the evaluation on Malay broadcast news transcription. To evaluate the performance of the proposed algorithm, a manually produced transcript from the spoken broadcast news corpus was used to evaluate the ASR results. The corpus used for this purpose are collectively referred to the MASS Malay speech corpus which consists of read speech corpus and broadcast news corpus [19]. The MASS Malay read speech corpus is utilized for training acoustic model. The speakers include Malay, Chinese and Indian speakers. The number of hours is approximately 122 hours. This dataset includes 199 speakers with approximately 58,420 utterances. Table 1 presents the details of the speech corpora.

**Table 1** The details of the speech corpora

| Feature | Number of speakers | Number of sentences |
|---|---|---|
| Malay | 74 | 21529 |
| Chinese | 112 | 32625 |
| Indian | 10 | 3301 |
| Other races | 3 | 965 |
| Total | 199 | 58420 |

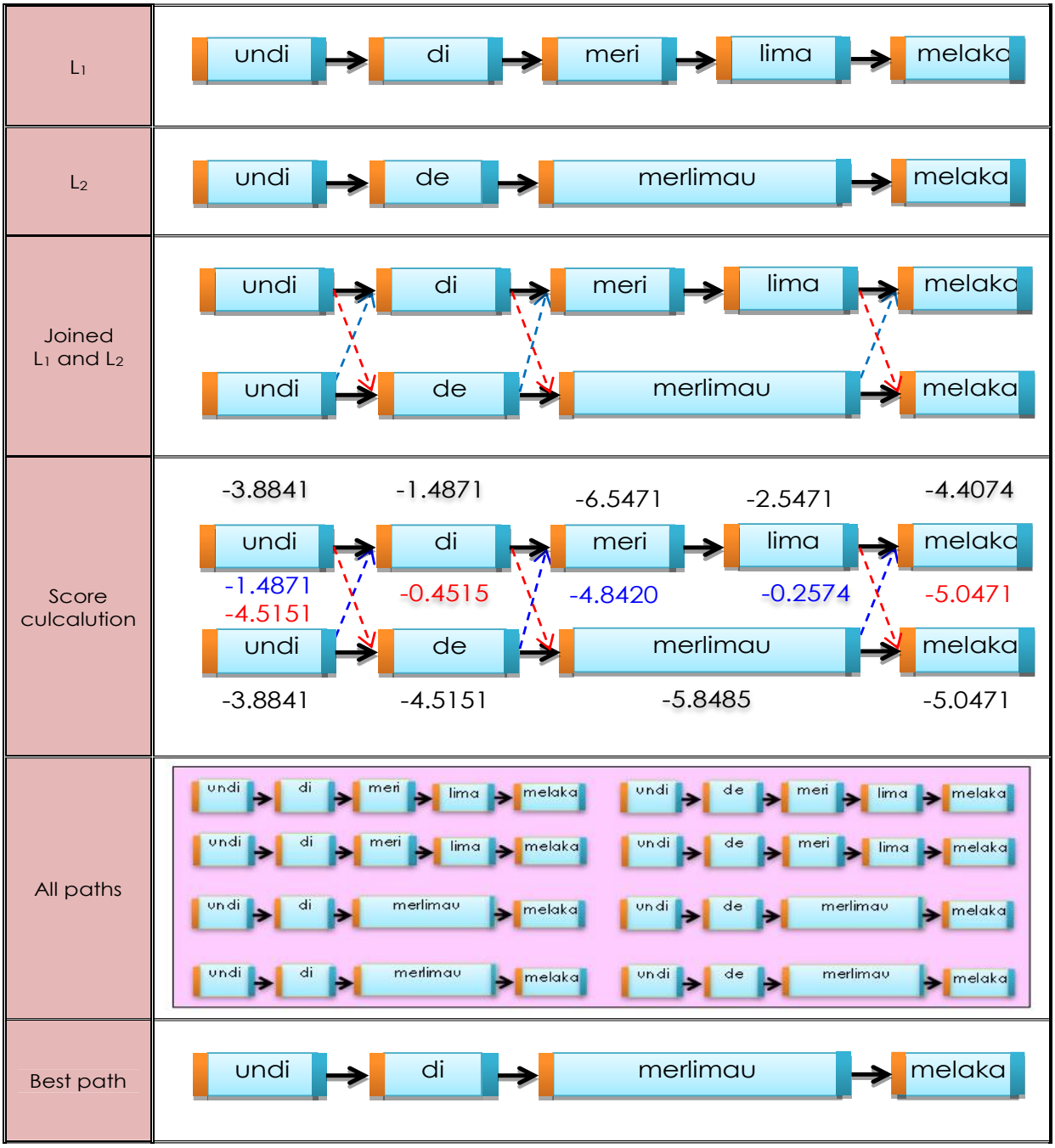



**Figure 7** An example of the two-pass rescoring system

All tests were carried out using part of the MASS broadcast news. The broadcast news contains speech from newscaster, reporter and interviewers in noisy environments. Broadcast news is recorded from local news channels in Malaysia (e.g. Television 3 (TV3) and Natseven Television (NTV7). The number of hours is approximately 10-hour, including different types of news such as local, political and sports news that are collected for different dates (roughly 18 dates) in March 2011. Also, the broadcast news included multiple speakers. The pronunciation model includes more than 75000 Malay words and variants.

## 5.0 TEXT CORPUS

To create trigram language models, the CMU Statistical Language Modeling toolkit (CMU SLM toolkit) was used [12]. A general domain trigram language model was built using the texts extracted from Malay news website [19]. The text corpus contains approximately 500 MB of text, from the year 1998 to 2011. Using Good Turing smoothing [20], a generic language model was created, and it consists of 42506 1-grams, 1763312 2-grams and 3053148 trigrams.

A domain specific language model was created from March 2011 broadcast news stories with using Good Turing smoothing. These news stories were collected automatically from the different Malay news websites and contain approximately 15000 webpages from different news domains. The size of news stories is 1.15 MB.

## 6.0 EXPERMINTAL RESULTS

This section presents the experiments to evaluate our proposed approaches. First, Validation method is explained in section 6.1. Then, the baseline of the ASR system that has been adapted will be described. Then, a system combination approach using ROVER will be presented. Next, the experimental performance of our proposed system will be examined. Finally, this section ends by discussing the experimental results.

### 6.1 ASR Validation

In the validation phase, each hypothesis that produced from the ASR is compared against a gold-standard (reference) that manually transcribed by humans [19] using alignment. To align hypotheses against reference, the NIST Scoring Toolkit (SCTK) package was used [21]. The Levenshtein edit distance is utilized for alignment.

The Levenshtein edit distance is the number of operations (OPR); insertions (I), substitutions (S) and deletions (D); required to transform a hypothesis into reference. For example, in Figure 8, to transform hypothesis transcription (HYP) into reference transcription (REF) requires one deletion ("*DI*"), one insertion ("*KE*"), and three substitution ("*AMBIL*" for "*ADIK*", "*LAKI*" for "*LELAKI*" and "*KARPA*" for "*KARPAL*"). The optimal alignment of two transcriptions is that which minimizes the Levenhshtein distance.

| | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| REF | mendiang | ADIK | LELAKI | KARPAL | jurubahasa | DI | mahkamah | tinggi | *** | pulau | pinang |
| HYP | mendiang | AMBIL | LAKI | KARPA | jurubahasa | *** | mahkamah | tinggi | KE | pulau | pinang |
| OPR | C | S | S | S | C | D | C | C | I | C | C |
| SCORES | #C (Correct)=6, #S (Substitutions)=3, #D (Deletions)=1, #I (Insertions)=1 | | | | | | | | | | |

**Figure 8** An example of the alignment approach

After alignment, sclite produces a variety of summary as well as detailed scoring reports. The evaluation matric that used for ASR system performance in sclite program is word error rate (WER). The traditional metric for evaluating the performance of ASR is the WER, which is defined as follows:

$$WER = 100 * \frac{\#S + \#I + \#D}{\text{No. of words in the correct transcript}} \quad (7)$$

The WER is 50% for the above example, which computes as follows:

WER= (3+1+1)/10*100=50%

### 6.2 Baseline Automatic Speech Recognition Result

After the acoustic model was trained as explained in section 4.0, we adapted the acoustic model using maximum a posteriori (MAP) and maximum likelihood linear regression (MLLR) algorithms [22]. We use a small dataset from the MASS corpus to adapt the acoustic model. The adaptation dataset consist of a list of sentences (10 sentences) by 13 numbers of speakers.

Then, the pronunciation model adaptation was carried out to predict pronunciation variants in the speech. The pronunciation variants were derived using decision trees and subsequently add them to the dictionary and retesting speech recognizer with the new pronunciation dictionary [22]. This system added 6732 variants to the dictionary. The adapted acoustic model and pronunciation model reduces the WER from 34.5% to 33.6%.

Language model interpolation is done by way of combination of many different language models [17]. WER is usually high when the language model of ASR system does not match the test domain. Therefore, language models interpolation was used

to improve the ASR performance by using a small specific testing data. Language model interpolation was applied on two trigram models. Generic language model and specific language model as was explained in section 3.1. Next, the generic language model was interpolated with the specific language model to create a new language model.

We use the CMU SLM toolkit [12] to create an interpolated language model from the generic and specific language models. In this tool, the expectation-maximization (EM) technique was utilized to compute interpolation weight by maximization of the likelihood of the data, as explained in Section 3.1. These weights were used to interpolate and combine the two language models. The best interpolation weight (the best lambda) that obtains from the CMU SLM toolkit was (0.688374, 0.311626). The ASR system applied with generic, specific and interpolated language models independently using the adapted acoustic model and pronunciation model give WER of 33.9%, 36.4% and 32.5% respectively. These results show that ASR performance with language model interpolation achieved better result compared with ASR performance with generic or specific language models. The ASR system using the interpolated language model is subsequently used as our baseline system in the following tests.

We are interested to know the performance of system combination approach that applies different language models. We evaluated system combination approached with ROVER from National Institute of Standards and Technology (NIST) Scoring Toolkit (SCTK) [21]. This system combines multiple ASR hypotheses to select the best scoring word sequence via a voting approach. Different scoring functions such as the word frequency, the word maximum confidence score (max CS) and the average confidence score (AVG CS), can be specified using the SCTK. The experiments for the approaches were performed as follows:

1. Two parallel ASR systems decode the speech utterances. Each ASR system uses the same pronunciation and acoustic models but a different language model. The Malay 1 automatic speech recognizer ($M_1$ ASR) uses the generic language model and the Malay 2 automatic speech recognizer ($M_2$ ASR) uses the interpolated language model to each produce a hypothetical word sequence. Each ASR produces a hypothesis sentence.
2. Combine these hypotheses into a single minimal cost word network WTN using a dynamic programming alignment approach.
3. The resulting network was re-evaluated and rescored via search approach to select the best word sequence with the highest number of votes.

The lowest WER was achieved using the CS; Nevertheless, this method achieves a 32.9% WER, which was worst that the baseline at 32.5%.

### 6.3 Proposed System Results

The same parallel ASR system that differs in the language models was used for decoding the input speech as described in the experiment with the ROVER system. The proposed post decoding system combination approach uses parallel automatic speech recognizers with two language models: the generic language model and the interpolated language model. First, a new language model created from in-domain data. Second, the generic language model interpolated with domain specific language model. The Malay 1 Automatic Speech Recognizer ($M_1$ ASR) with generic LM and Malay 2 Automatic Speech Recognizer ($M_2$ ASR) with interpolation language model produce a word sequence that contains only the most probable word sequence given the observation. In the rescoring process, the word sequences are joined to form a single sequence of words to be considered as a hypothesis. The goal of this stage of the experiment is to select the correct recognizer for each word in the utterance. The experiments for the approaches were performed as follows.

- Use Malay 1 Automatic Speech Recognizer ($M_1$ ASR) and Malay 2 Automatic Speech Recognizer ($M_2$ ASR) to recognize each utterance as word sequences.
- Merge the two word sequences from the previous step based on the acoustic weight, language weight, start frame, and end frame.
- Apply the rescoring process, where the word sequences are joined to form a single sequence of words to be considered as a hypothesis.

The proposed post decoding system combination approach with different weights ($w$) applied on the hypothesis sentence produced using ASR with domain specific language model and on hypothesis sentence produced using ASR with generic language model ($1$-$w$) shows improvement in the WER. Table 2 shows the proposed post decoding system combination performance with different weights. We tested our hypotheses using different interpolation weights ($0 < \text{weight} < 1$). The experimental result of applying our proposed framework with different weights for the language models shows that at weight 0.9, the WER is the lowest at 30.6% with 2.9% reduction. See Table 2.

**Table 2** Proposed system performance with different weights

| Weight | 0.1 | 0.2 | 0.3 | 0.4 | 0.5 | 0.6 | 0.7 | 0.8 | **0.9** |
|---|---|---|---|---|---|---|---|---|---|
| **WER%** | 32.6 | 32.4 | 32.1 | 31.9 | 31.7 | 31.4 | 31.2 | 30.9 | **30.6** |

In this paper, the acoustic model (AM) adaptation, the pronunciation model (PM) adaptation, language model interpolation and a conventional voting scheme, named ROVER and are used as baselines. Finally the performances of the baseline and the proposed system for the same dataset are compared and discussed. Table 3 compares the baselines and the proposed system performances for the same dataset.

**Table 3** The baselines and the proposed system performances

| Approach | WER% |
|---|---|
| ASR system with acoustic model adaptation | 33.9 |
| ASR system with PM adaptation | 33.6 |
| ASR system with LM interpolated model | 32.5 |
| ROVER system | 32.9 |
| Proposed Post Decoding System Combination | 30.6 |

The proposed system achieved an overall 30.6% WER with 1.9% WER reduction when tested on Malay broadcast news.

### 6.4 Discussion

The study propose system combination approaches using parallel ASR systems with two different language models that generate two hypotheses to reduce the WER in the ASR system. The proposed approach achieves a lower WER than baseline ASR at 32.5%. In addition, the proposed systems achieve better results compared with the popular approach of multi-hypothesis combination system, i.e., ROVER system (baseline 2); the WER was 32.9%. A total improvement to 30.6% WER was achieved using proposed system.

The ASR system combination reduces the WER to 30.6% for the experimental data with proposed method. The main aim of the with proposed approach is to find the best word sequence for each utterance based on timing information by combining the hypotheses produced from specific and generic language models. The main factor that impact on the WER in proposed method is the word sequence. Thus, the main aim of the proposed approach is to find the best word sequence for each utterance based on timing information between the fusion of the specific and generic language models. The total relative improvement in the word sequences was 3.93%.

## 7.0 CONCLUSIONS

To recap, the results prove that proposed method were able to improve the Malay broadcast news. The improvements resulting from integrating two language models into an interpolated language model using the proposed systems were compared with the adapted model (acoustic model adaptation, pronunciation model adaptation and language model interpolation) and with the post-decoding system combination (ROVER system). We examine a post decoding system combination approach that combines the hypotheses of two ASR systems using different language model. The main idea of the post decoding system combination approach is to take the advantage of multiple ASR hypotheses to achieve accurate ASR hypothesis results. This advantage is based on the time and score information provided by the 1-best lattice, especially the language model score, to find the most likely words to select them as the final output. The experiment results show that the technique reduce the baseline WER and performed better than ROVER system. The average relative WER improvements derived from using the proposed method were 11.3%.

For future work, several improvements to the system were identified. For instance:

1. Apply acoustic model trained using broadcast news.
2. Apply the proposed algorithms to dialog, conversation, interviews and meetings.
3. Test the proposed algorithms in other languages, such as English and Arabic languages.

## Acknowledgement

This work is supported by USM University Grant 1001/PKOMP/817068.

## References

[1] Grangier, David, & Vinciarelli, Alessandro. 2005. Effect of Segmentation Method on Video Retrieval Performance. In *IEEE International Conference on Multimedia and Expo (ICME-05)*, IEEE, *Amsterdam, The Netherlands.* 5-8.

[2] Wu, *Chung-Hsien, & Hsieh, Chia-Hsin*. 2009. Story Segmentation and Topic Classification of Broadcast News Via a Topic-Based Segmental Model and a Genetic Algorithm. *IEEE Transactions on Audio, Speech, and Language Processing.* 17(8): 1612-1623.

[3] Lu, Mi Mi, Xie, Lei, Fu, Zhong Hua , Jiang, Dong Mei , & Zhang, Yan Ning. 2010. Multi-Modal Feature Integration for Story Boundary Detection in Broadcast News. In *7th International Symposium on Chinese Spoken Language processing (ISCSLP), IEEE, Taiwan*. 420-425.

[4] Lojka, M., & Juhar, J. 2014. Hypothesis Combination for Slovak Dictation Speech Recognition. In *56th International*

*Symposium Electronics in Marine (ELMAR), IEEE, Zadar, Croatia.*1-4.

[5] Ellis, Daniel P. W. 2000. Stream Combination Before and/or After the Acoustic Model. In *Proceedings of the International Conference on Acoustics, Speech and Signal Processing*, Istanbul, Turkey. 1635-1638.

[6] Hoffmeister, Björn, Klein, Tobias, Schlüter, Ralf, & Ney, Hermann. 2006. Frame based System Combination and a Comparison with Weighted ROVER and CNC. In *International Conference on Spoken Language Processing, Interspeech*, Pittsburgh, PA, USA. 537-540.

[7] Hoffmeister, Björn, Schlüter, Ralf, & Ney, Hermann. 2008. iCNC and iROVER: The Limits of Improving System Combination with Classification? *In the 9th Annual Conference of the International Speech Communication Association*, Interspeech, Brisbane, Australia. 232-235.

[8] Chen, I-Fan, & Lee, Lin-Shan. 2006. A New Framework for System Combination Based on Integrated Hypothesis Space. In *Proceedings of the International Conference on Spoken Language Processing (ICSLP)*, Pittburgh, Pensylvania, USA. 533-536.

[9] Fiscus, Jonathan G. 1997. A Post-Processing System to Yield Reduced Word Error Rates: Recognizer Output Voting Error Reduction (ROVER). In *IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU),* Santa Barbara, CA, USA. 347-352.

[10] Schwenk, Holger, & Gauvain, Jean-Luc. 2000. Combining Multiple Speech Recognizers Using Voting and Language Model Information. In *International Conference on Spoken Language Processing, (ICSLP 2000),* Beijing, China. 915-918.

[11] Stolcke, A., Bratt, H., Butzberger, J., Franco, H., Gadde, V. R., Plauché, M., & Zheng, J. 2000. The SRI March 2000 Hub-5 Conversational Speech Transcription System. In *Proceedings of the NIST Speech Transcription Workshop*.

[12] Clarkson, Philip, & Rosenfeld, Rosenfeld. 1997. Statistical Language Modeling Using the CMU-Cambridge Toolkit. In *5th European Conference on Speech Communication and Technology*, Rhodes, Greece. 2707-2710.

[13] Goel, Vaibhava, Kumar, Shankar, & Byrne, William. 2000. Segmental Minimum Bayes-risk ASR Voting Strategies. In *Proceedings of the International Conference on Spoken Language Processing (ICSLP)*, Beijing, China. 139-142.

[14] Hillard, Dustin, Hoffmeister, Björn, Ostendorf, Mari, Schlüter, Ralf, & Ney, Hermann. 2007. iROVER: Improving System Combination with Classification. In *Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics; Companion Volume, Short Papers* , Association for Computational Linguistics, Rochester, New York, USA. 65-68.

[15] Zhang, R., & Rudnicky, A. 2006. Investigations of Issues for Using Multiple Acoustic Models to Improve Continuous Speech Recognition. In *International Conference on Spoken Language Processing (ICSLP),* Pittsburgh, PA, USA. 529-533.

[16] Ahmed, Basem A. 2014. Automatic Speech Recognition for MultiLingual Speakers. Ph.D. thesis, Universiti Sains Malaysia, Malaysia.

[17] Brychcn, Tomas. 2012. *Unsupervised Methods for Language Modeling*. Ph.D. thesis, University of West Bohemia in Pilsen, Czech Republic.

[18] Jayalakshmi, T., & Santhakumaran, D. A. 2011. Statistical Normalization and Back Propagation for Classification. *International Journal of Computer Theory and Engineering.* 3(1): 1793-8201.

[19] Tan, Tien Ping, Haizhou, L. L., Kong, Tang Enya, & Xiong, Xiao. 2009. Mass: A Malay Language LVCSR Corpus Resource. In *International Conference on Speech Database and Assessments, 2009 Oriental COCOSDA*, IEEE, Urumqi. 25-30.

[20] Good, Irving J. 1953. The Population Frequencies of Species and the Estimation of Population Parameters. *Biometrika.* 40(3-4): 327-264.

[21] *The NIST Scoring Toolkit (SCTK). Available: http://www1.icsi.berkeley.edu/Speech/docs/sctk-1.2/sctk.htm*

[22] Tan, Tien Ping. 2008. *Automatic Speech Recognition for Non-Native Speakers.* Ph.D. thesis, Université Joseph Fourier.