A study on the effects of limited training data for English, Spanish and Indonesian keyword spotting

K. Thambiratnam, T. Martin and S. Sridharan

Speech and Audio Research Laboratory
Queensland University of Technology
GPO Box 2434, Brisbane, Australia 4001
[k.thambiratnam,tl.martin,s.sridharan]@qut.edu.au

Abstract

This paper reports on experiments to quantify the benefits of large training databases for non-English HMM-based keyword spotting. The research was motivated by the lack of such databases for many non-English languages, and aims to determine if the significant cost and delay of creating these databases justifies the gains in keyword spotting performance. HMM-based keyword spotting experiments performed for English, Spanish and Indonesian found that although some gains in performance can be obtained through increased training database size, the magnitude of these gains may not necessarily justify the effort and incurred delay of constructing such databases. This has ramifications for the immediate development and deployment of non-English keyword spotting systems.

1. Introduction

With the recent increase in global security awareness, non-English speech processing has emerged as a major topic of interest. One problem that has hindered the development of robust non-English keyword spotters is the lack of large transcribed non-English speech databases. This paper reports on experiments to quantify the benefits of large training databases for non-English keyword spotting. Specifically it aims to determine if the significant cost of collecting and transcribing large non-English databases justifies the gains in keyword spotting performance. This has ramifications for the immediate development and deployment of non-English keyword spotting systems.

A study on the effect of training database size reported in (Moore 2003) demonstrated the merits of large training databases for speech transcription. This study revealed that gains in word error rate were significant when comparing systems trained on a few hours of speech with systems trained on tens and hundreds of hours of speech. Although some of the word error rate gains were from more robust acoustic models, a major component was also sourced from more robustly trained language models.

In keyword spotting, language models do not play as significant a role. Specifically, HMM-based keyword spotting (Rohlicek 1995) and speech background model keyword verification (Wilpon, Rabiner, Lee, and Goldman 1990) do not require language models at all. In fact these two algorithms perform a much simpler task than speech transcription. For example single keyword spotting is essentially a two-class discrimination task relying completely on acoustic models. In view of the reduced complexity of the keyword spotting task, it is plausible that keyword spotting performance is less sensitive to training database size.

Keyword spotting and verification experiments were performed for English, Spanish and Indonesian using a variety of training database sizes. Experiments for Spanish and Indonesian were only done on smaller sized databases as there was significantly less transcribed data available. Trends in performance across training database size were examined, as well as the effects of different model architectures (eg. monophones versus triphones). Finally predictions for expected performance of the Indonesian keyword spotter trained on a larger database were made based on trends observed in English and Spanish experiments.

2. Background

Hidden Markhov Model (HMM) based speech recognition provides a convenient framework for keyword spotting. The techniques for training such systems are well established and training methods can remain independent of the target language. A two stage approach is used in the reported evaluations. First, a HMM-based keyword spotter is used to generate a set of candidate keyword occurrences. A subsequent speech background model keyword verification stage is then used to prune false alarms (FAs).

2.1. HMM-based keyword spotting

A keyword spotter is used to postulate candidate occurrences of a target keyword in continuous speech. HMM-based keyword spotting (HMMKS) uses a speech recogniser to locate these candidate occurrences. All non-target-keywords in the target domain's vocabulary are represented by a 'non-keyword' word. An open word loop recognition network is then used to locate candidate keyword occurrences. The grammar to perform HMMKS is given by the Extended Backus-Naur Form grammar:

$$\begin{aligned} \mathit{Keywords} &::= \mathit{keyword}1|\ldots|\mathit{keywordN} && (1)\\ \mathit{Grammar} &::= (\mathit{Keywords}|\mathit{nonkeyword})+ && (2)\end{aligned}$$

Recognition using this grammar generates a time-marked sequence of keyword and non-keyword tokens for a given observation sequence.

Ideally the non-keyword model should model all non-target-keywords in the target domain's vocabulary. However this is not only complex but computationally expensive and hence a plethora of non-keyword model approximations have been proposed in literature. These include anti-syllable models (Xin and Wang 2001), a uniform distribution (Silaghi and Bourlard 2000) and a speech background model (Wilpon, Rabiner, Lee, and Goldman 1990). For the experiments reported in this paper, the speech background model (SBM) described in (Wilpon et al. 1990) was selected as the non-keyword model of choice because of it's prevalent use in many other areas of speech research.

The algorithm for HMMKS using an SBM (HMMKS-SBM) is:

1. Given a set of target keywords, create a recognition network using the grammar in equation 2
2. For each utterance, use a speech recogniser and the constructed recognition network to generate a sequence of keyword/non-keyword tokens
3. Select all keyword tokens in the recogniser output sequence and label them as candidate keyword occurrences
4. The candidate occurrences are passed on to a subsequent keyword verification stage to cull FAs

2.2. Speech background model keyword verification

Keyword verification algorithms are used to reduce the number of FAs output by a preceding keyword spotting stage. Typically such algorithms derive a confidence score for each candidate keyword occurrence and then accept or reject the candidate by thresholding. In Log-Likelihood Ratio (LLR) based keyword verification, the keyword confidence scoring metric takes the form:

$$C = log(p(X|\lambda_{keyword})) - log(p(X|\lambda_{nonkeyword})) \qquad (3)$$

where X is the sequence of observations corresponding to the candidate to be verified, $\lambda_{keyword}$ is the acoustic model for the target keyword (eg. concatenated monophones or triphones) and $\lambda_{nonkeyword}$ is the acoustic model for the non-keyword against which the target word is scored. The non-keyword model is analogous to the non-keyword model used in HMMKS.

Verification performance can vary dramatically depending on the choice of non-keyword model. For example cohort word non-keyword models were shown to yield better performance than Gaussian Mixture Model non-keyword models in (Thambiratnam and Sridharan 2003). For the experiments reported in this paper, the SBM used in the HMMKS stage was also used as the non-keyword model for keyword verification to provide consistency between the spotting and verification stages. The LLR-based confidence score for a candidate keyword occurrence using an SBM is then given by:

$$C = log(p(X|\lambda_{keyword})) - log(p(X|\lambda_{SBM})) \qquad (4)$$

Given the confidence score formulation in equation 4, the algorithm for speech background model keyword verification (SBMKV) is:

1. For each candidate, calculate the SBMKV confidence score given by equation 4
2. Apply thresholding using the SBMKV confidence score to accept/reject candidates

3. Experiment Setup

Training and evaluation speech were taken from the Switchboard 1 English telephone speech corpus, the Callhome Spanish telephone speech corpus and the OGI Multilingual Indonesian telephone speech corpus. For each language all utterances containing out-of-vocabulary words were removed. This gave a total of approximately 165 hours of English data, 10.2 hours of Spanish data, and 3.5 hours of Indonesian data. Due to the limited amount of data available for the non-English languages, we designated only 40 minutes of data from each language set as evaluation data while the remaining data was used for training.

All data was parameterised using Perceptual Linear Prediction (PLP) coefficient feature extraction. Utterance based cepstral mean subtraction (CMS) was applied to reduce the effects of channel/speaker mismatch.

3.1. Training data sets

Reduced size training sets were generated for English and Spanish by randomly selecting utterances from the full sized training sets. Since there were only 2.8 hours of data for Indonesian, it was decided that the smallest training set size for the other languages would be of a comparable size. However, as the size of phoneset differed between languages (44 for English, 28 for Spanish and 28 for Indonesian), the average number of hours of speech per phone instead of the total number of hours of speech was kept constant across the reduced size training data sets. This resulted in reduced size training sets of 4.2 hours for English, 2.8 hours for Spanish and 2.8 hours for Indonesian, an average of approximately 0.1 hours per phone (h/phone) for each data set.

An intermediate sized English training database was also created to facilitate comparative experiments between English and the full-sized Spanish training database. As before, the average number of hours of speech per phone was kept consistent between the two languages. This gave an intermediate sized English training database of 15.4 hours, approximately 0.35 h/phone.

To avoid confusion, the codes in table 1 are used when referring to the individual training data sets. The S1 training sets correspond to the 0.1 h/phone training data sets and exist for all three languages. The S2 training sets correspond to the 0.35 h/phone training data sets and only exist for English and Spanish. Finally the S3E set corresponds to the full sized English training data set and was included to provide insight into spotting and verification performance for systems trained using very large databases.

3.2. Model architectures

Three HMM phone model architectures were trained for each training data set: 16-mixture monophones, 32-mixture monophones and 16-mixture triphones. It was anticipated that the triphone architecture would provide the greatest performance when using the large training data sets but would have reduced performance for smaller training data sets due to data sparsity issues. The 16-mixture monophone and 32-mixture monophone architectures were included to address these data sparsity issues. Finally a 256-mixture

Code	Language	Hours of speech	Hours per phone
S1E	English	4.15	0.095
S1S	Spanish	2.82	0.10
S1I	Indonesian	2.78	0.099
S2E	English	15.4	0.35
S2S	Spanish	9.59	0.34
S3E	English	164.05	3.73

Table 1: *Summary of training data sets*

GMM SBM was trained for each training database for use with the HMMKS-SBM and SBMKV algorithms.

To facilitate ease of reference to the numerous model sets, the label M16 is used when referring to 16-mixture monophone models, M32 for 32-mixture monophone models, T16 for 16-mixture triphone models, and G256 for SBM models. Furthermore, when referring to a model trained on a specific training set, the name of the training set is appended to the model label. Hence, a 16-mixture triphone model set trained on the S2S training set is referred to as the T16S2S model set whereas the SBM trained on the S1I set is referred to as the G256S1I model set.

3.3. Evaluation procedure

The evaluation data sets consisted of approximately 40 minutes for each language. It was not possible to use a larger evaluation set because of the limited amount of data available for Indonesian and Spanish.

For English and Spanish, 180 unique words of medium length (6 phones) were randomly selected for each language and designated as the evaluation query word set. In contrast only 150 words were selected for Indonesian as there were only 153 unique medium-length words in the Indonesian evaluation set. Table 2 summarises each evaluation set. The ’instances of query words in eval data’ number corresponds to the number of instances of the words in the query word set that occur in the evaluation data ie. the total number of hits required to obtain a miss rate of 0%.

Code	Language	Mins of speech	Num query words	Instances query words in eval data
EE	English	43.62	180	298
ES	Spanish	39.60	180	353
EI	Indonesian	41.40	150	349

Table 2: *Summary of evaluation data sets.*

Experiments were performed to evaluate the effect of training database size on spotting and verification performance for each of the three target languages. Additionally, the experiments were repeated using the various model architectures described in section 3.2. The evaluation procedure used was:

1. Perform keyword spotting using HMMKS-SBM for each word in the evaluation query word set on each utterance in the evaluation speech set.
2. Calculate miss and FA/keyword/hour rates. These results were termed the raw spotting miss rate and raw spotting FA/kwd-hr rate.
3. Perform keyword verification using SBMKV on the output of the keyword spotting stage.
4. Calculate miss, FA, and equal error rates (EERs) for the SBMKV output over a range of acceptance thresholds. These results were termed the post verification miss probabilities, post verification FA probabilities and the post verification EERs.

4. Results

4.1. English and Spanish raw keyword spotting experiments

Experiments were first performed to evaluate raw spotting miss and FA rates for the various English and Spanish models. Of particular interest was the effect of training database size on raw spotting miss rate, as this gives a lower bound on the achievable miss rate for a successive keyword verification stage. Each model set was evaluated on the appropriate evaluation set for the language and using the SBM trained on the same data set. Table 3 shows the results of these experiments.

Model	Miss rate	FA/kw-hr
M16S1E	4.0	929.2
M32S1E	2.3	1214.4
T16S1E	5.7	368.7
M16S2E	2.7	965.7
M32S2E	2.0	1361.1
T16S2E	5.0	307.0
M16S3E	3.7	945.6
M32S3E	2.3	1374.6
T16S3E	1.0	297.1
M16S1S	7.6	818.1
M32S1S	4.5	1110.8
T16S1S	11.9	305.5
M16S2S	6.2	918.4
M32S2S	3.7	1321.5
T16S2S	10.8	316.3

Table 3: *Raw spotting rates for various model sets and training database sizes*

A number of observations can be made regarding the raw spotting rates. Of note is that the Spanish miss rates were much higher than the English miss rates. One explanation for this poorer performance is that the average utterance duration for the Spanish data was shorter than that of the English data. Since CMS was being used, the shorter utterance length could lead to poorer estimates for the cepstral mean, and therefore a decrease in recognition performance. An equally likely explanation is that the Spanish data was simply more difficult to recognise due to factors such as increased speaking rate and background noise.

The results demonstrate that in most cases increased training database size resulted in decreased miss rates and increased FA/kwd-hr. A decrease in miss rate is beneficial as this reduces the lower bound for the minimum achievable miss rate for a subsequent keyword verification stage. Final post-verification FA may not necessarily be dramatically impacted by increased FA/kwd-hr at this stage if the verifier is able to prune the extra FAs. Interestingly though, the absolute gains in miss rate were not particularly large. Apart from the gain observed for the T16S3E system, the other gains were below 2%, and in most cases below 1%. This implies that FKSM miss rate is not dramatically affected by training database size.

An unexpected result was that the English monophone S3 models resulted in increased miss rates compared to the corresponding S2 models. This is in opposition to trends observed for the other experiments. A likely explanation for this result is that the monophone architectures were too simple to train compact discriminative models using the larger S3 database.

Performance gains also varied with model architecture. While the M16 and M32 architectures outperformed the T16 architectures for the smaller S1 and S2 training data sets, the converse was observed for the S3 experiments. This suggests that there was insufficient data to train robust triphone models for smaller training data sets and too much data to produce robust monophone models for the large training data sets.

The triphone architectures also provided significantly lower FA/kw-hr rate than the monophone architectures for all training data set sizes. One may argue that this is simply a trade-off in performance - a lower FA/kw-hr result in exchange for a higher miss rate. This appears to be the case for the Spanish experiments. However, in the English experiments, both miss rate and FA/kw-hr rates decreased as training data size was increased. From these limited set of experiments, it is not possible to determine whether the triphone architecture truly provides an increase in both rates or simply a trade off between the two measures.

Overall, increased training database size does yield improved performance in miss rate, though the gains are not dramatic unless very large database sizes are used. For S1 and S2 sized databases, the monophone architectures yielded more favorable miss rates at the expense of significantly higher FA/kw-hr rates.

4.2. English and Spanish post keyword verification experiments

Joint HMMKS-SBM/SBMKV performance was evaluated for the various English and Spanish training databases and model architectures. The aim of these experiments was to determine the effect of training database size on the final keyword spotting performance for a combined HMMKS-SBM/SBMKV system, as opposed to the effect on isolated SBMKV. This is because in practice the same data sets would be used when training models for the spotting and verification stages. Hence HMMKS-SBM followed by SBMKV was performed and the final miss and FA probabilities at a range of acceptance thresholds were measured.

Table 4 shows the EERs after SBMKV for the various English and Spanish model types. Figures 1, 2 and 3 show the detection error trade-off plots for the T16, M16 and M32 experiments respectively. A number of trends can be seen in these results.

Model	EER	Model	EER
	rate		rate
M16S1E	22.2	M16S1S	25.8
M32S1E	19.1	M32S1S	24.4
T16S1E	18.1	T16S1S	28.7
M16S2E	19.8	M16S2S	25.2
M32S2E	17.8	M32S2S	22.6
T16S2E	17.8	T16S2S	26.9
M16S3E	20.5		
M32S3E	18.5		
T16S3E	13.0		

Table 4: *Equal error rates after SBMKV for various model sets and training database sizes*

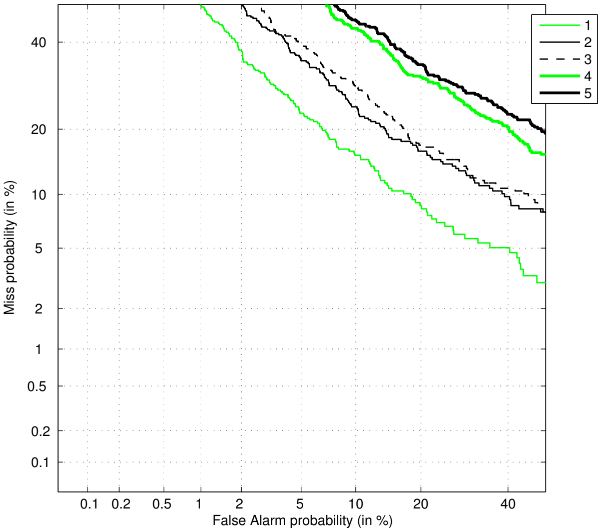

Figure 1: *Detection error trade-off for T16 SBMKV experiments. 1=T16S3E, 2=T16S2E, 3=T16S1E, 4=T16S2S, 5=T16S1S*

Of note is the gain in performance between the S1 and S2 systems given a fixed model architecture. In most cases, increasing the amount of training data from the S1 to S2 database size resulted in absolute gains of approximately 1-2% in EER. Further increasing the database size as done in the S3 experiments resulted in gains for the triphone system only (4.8% absolute).

This is a positive result, indicating that the relatively small increase in training database size between S1 and S2 provided a tangible gain in performance. Furthermore, the fact that a significantly larger training database only yielded a 4.8% absolute gain for the T16S3E experiment suggests that returns diminish with increases in training database size.

This observation has important ramifications for the development and deployment of keyword spotting systems. It indicates that HMMKS-SBM/SBMKV systems trained on relatively small databases are able to achieve performances

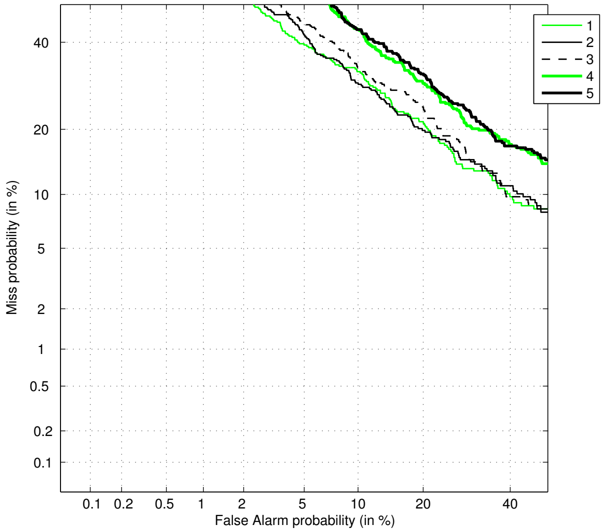

Figure 2: *Detection error trade-off for M16 SBMKV experiments. 1=M16S3E, 2=M16S2E, 3=M16S1E, 4=M16S2S, 5=M16S1S*

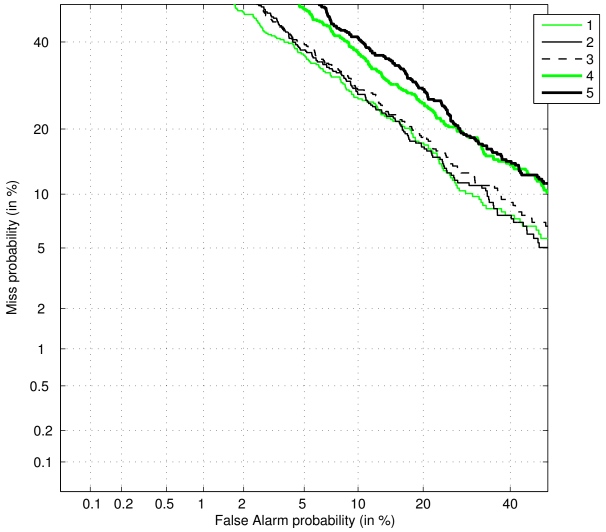

Figure 3: *Detection error trade-off for M32 SBMKV experiments. 1=M32S3E, 2=M32S2E, 3=M32S1E, 4=M32S2S, 5=M32S1S*

well within an order of magnitude of systems trained using significantly larger databases. Depending on the target application, this loss in performance may be an acceptable trade-off for the time and monetary costs of obtaining larger databases.

Another observation is the difference in EER gains observed for English triphone systems over English monophone systems compared to those observed for the equivalent Spanish systems. In all cases, the English triphone systems markedly outperformed the monophone systems, whereas for Spanish, the triphone systems yielded considerably lower EERs compared to the monophone systems. Further analysis of the data revealed that for the S1S and S2S evaluations, the M32 systems outperformed the performance of the T16 systems at all operating points (see figure 4).

One possible explanation for the disparity in performance gains between the English and Spanish triphone systems is the decision tree clustering process used during triphone training. The question set used for the English decision tree clustering process was a well established and well tested question set, whereas the Spanish question set was a relatively new question set constructed for this particular set of experiments. Although much care was taken in building the Spanish question set and in removing any errors, it is possible that the nature of the phonetic questions asked, though relevant and applicable to English, were not suitable for Spanish decision tree clustering.

In summary, the experiments demonstrate that although some gains in performance were achieved using larger training databases, the magnitude of these gains were not dramatic and may not justify the costs of obtaining such databases. For smaller-sized databases, the M32 architecture resulted in more robust performance for Spanish keyword spotting, though this may be due to issues with the triphone training procedures for Spanish.

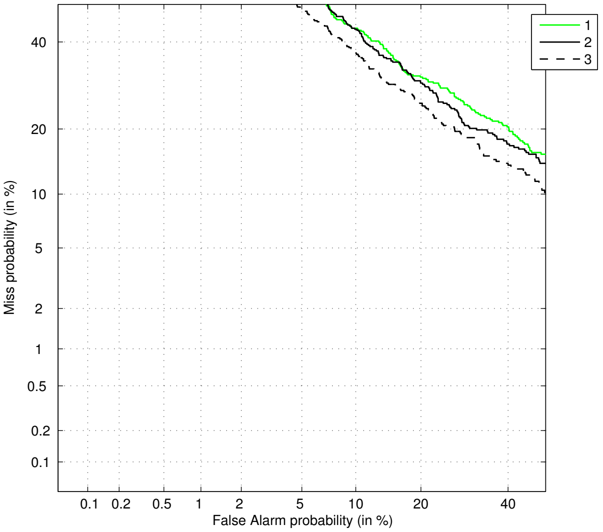

Figure 4: *Detection error trade-off for S2S SBMKV experiments. 1=T16S2S, 2=M16S2S, 3=M32S2S*

4.3. Indonesian Keyword spotting and verification experiments

Given the results and trends observed in the English and Spanish experiments, evaluations were performed using the small amount of available Indonesian data to obtain baseline keyword spotting performance. Table 5 and figure 5 show the results of these experiments.

Model	Raw spot miss rate	Raw spot FA/kw-hr	Post-verifier EER
M16S1I	3.4	393.2	22.0
M32S1I	3.0	439.1	21.0
T16S1I	3.4	394.8	22.0

Table 5: *Raw spotting and post verification results for S1I experiments*

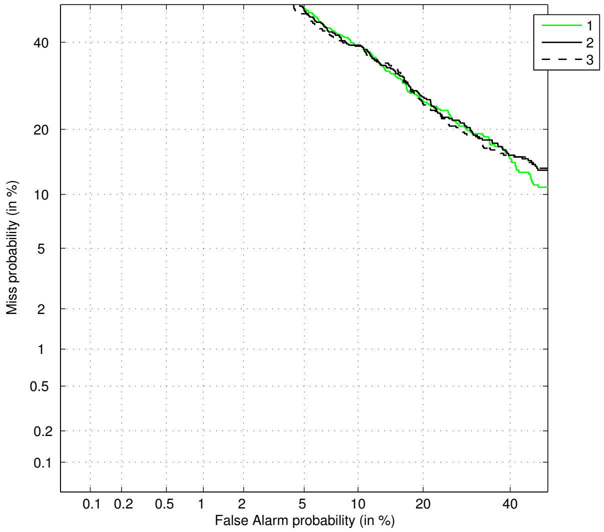

Figure 5: *Detection error trade-off plot for S1I SBMKV experiments. 1=T16S1I, 2=M16S1I, 3=M32S1I*

Raw spotting performance results were not as diverse as those observed for English and Spanish - all models yielded similar miss rates and comparable FA/kw-hr rates. In contrast, the trends for post-verifier EER were similar to that observed for Spanish, with the M32 architecture yielding the best EER performance and in fact the best performance at most other operating points. Ultimately though, as demonstrated by figure 5, the post verification performance for all model types were very close, being within 1% absolute in most cases.

Given the consistent 1-2% EER gain observed when increasing from S1 to S2 sized training data sets for the English and Spanish experiments, it is reasonable to postulate that similar gains in EER would be observed in Indonesian. However, any such extrapolations would have a low degree of confidence since there are many language-specific factors that could increase or decrease these gains. All things being equal though, it would not be unreasonable to expect a similar 1-2% gain in EER for a S2-sized training database.

Extrapolations regarding expected EER gain for a S3-sized database would have an even lower degree of confidence than those for the S2-sized database since consistent trends were not observed in the S3E experiments across the various model types. Difficulties of extrapolation are further compounded by the fact that the trends in triphone performance observed for English were different to those observed for Spanish, potentially due to problems with the Spanish triphone training methods. Nevertheless it is reasonable to assume that an Indonesian S3-trained triphone system would not outperform a T16S3E system in light of the poorer Indonesian S1 performance. Therefore at the very best, a properly trained Indonesian S3-trained triphone system would achieve an EER equal to the T16S3E system (12.5%). More realistically though, one would expect a T16S3I EER in the vicinity of 14-16% (1-2% S2 EER gain plus 4-5% S3 EER gain) given the 4.8% EER gain observed for the T16S3E system over the T16S2E system.

5. Conclusions

The experiments demonstrate that the development and deployment of a non-English HMMKS-SBM/SBMKV using small training databases is realistic and not overly suboptimal. Though some gains can be obtained through increased training database size, the magnitude of gains (eg. the very best being 4.8% for a triphone English system) may not necessarily justify the effort of collecting and transcribing a significantly larger training database. This is particularly relevant for non-English target domains where data collection and transcription is markedly more difficult and costly. For the present, non-English keyword spotting systems can feasibly be developed with small training databases and still achieve performances close to that of a system trained using a very large database.

In addition, the experiments show that a M32 system is more robust than a T16 system for non-English keyword HMMKS-SBM/SBMKV keyword spotting using smaller sized training databases. However, this may be a result of inappropriate non-English triphone training procedures since the English 16-mixture triphone system did yield better performance than corresponding 32-mixture monophone system for the smaller sized databases.

Low confidence extrapolations were also made regarding expected equal error rate gains for an Indonesian HMMKS-SBM/SBMKV keyword spotting system trained on a large database. A system trained on 2.8 hours of training data yielded an EER of 21.0% using a 32-mixture monophone model set. Trends seen in English and Spanish imply an Indonesian HMMKS-SBM/SBMKV EER gain of 1-2% using a 9.6 hour database and a further gain of 4-5% using a significantly larger training database.

References

Moore, R. (2003). A comparison of the data requirements of automatic speech recognition systems and human listeners. In *Proceedings of Eurospeech 2003*, Geneva, Switzerland.

Rohlicek, J. R. (1995). *Modern methods of Speech Processing*, Chapter Word Spotting, pp. 136–140. Kluwer Academic Publishers.

Silaghi, M. and H. Bourlard (2000). A new keyword spotting approach based on iterative dynamic programming. In *IEEE Internation Conference on Acoustics, Speech and Signal Processing 2000*.

Thambiratnam, K. and S. Sridharan (2003). Isolated word verification using cohort word-level verification. In *Proceedings of Eurospeech 2003*, Geneva, Switzerland.

Wilpon, J. G., L. R. Rabiner, C. H. Lee, and E. R. Goldman (1990). Automatic recognition of keywords in unconstrained speech using hidden markov models. *IEEE Transactions on Acoustics, Speech and Signal Processing 38*, 1870–1878.

Xin, L. and B. Wang (2001). Utterance verification for spontaneous mandarin speech keyword spotting. In *Proceedings ICII 2001*, Beijing.