THREE APPROACHES TO MULTILINGUAL PHONE RECOGNITION

Eddie Wong and Sridha Sridharan

Speech Research Lab, RCSAVT
School of Electrical and Electronic Systems Engineering
Queensland University of Technology
2 George St, Brisbane, Q4001, Australia
{ee.wong, s.sridharan}@qut.edu.au

ABSTRACT

This paper investigates and compares three different approaches of multilingual phone recognition (MPR). Two types of MPR approach are defined according to the Language Identification (LID) process of the system: Explicit-LID where language identification is mandatory, and Implicit-LID where LID is an integrated part of the MPR process. The OGI-TS database is employed to perform the isolated and continuous MPR experiments. Three of the world's most spoken languages; English, Mandarin and Spanish are selected as the target languages for the system. Experimental results indicate that different MPR approaches should be employed for different applications according to the degree of LID accuracy that can be achieved from the input test utterance. If high LID accuracy is achievable, the MPR approach that depends on LID can obtain better performance. Conversely, the Implicit-LID MPR approach is more appropriate.

1. INTRODUCTION

The ability to process speech in multiple languages by a *single* speech recognition system has become increasingly desirable due to the trend of globalisation and the popularity pervasive of the Internet. This multilingual feature not only extended the usability of the system, but also allows it to process a larger range of speech data.

One of the more popular approaches to perform multilingual speech recognition is the utilisation of a multilingual phone set. These multilingual phones are usually created by merging phones across the target languages that are acoustically similar in an attempt to obtain a minimal phone set that covers all the sounds that exist in all the target languages [1,2,3]. One important application of this approach is that the multilingual phone set can be adapted to recognise speech of an unknown language with no or limited adaptation speech data. Unfortunately, performance of the system employing this approach was not comparable to its monolingual counter part. Therefore, more investigation on this and other multilingual approaches are necessary.

In this paper, research on multilingual speech recognition is focused at the phone level using the OGI-TS corpus [4]. Three of the most spoken languages in the world; English, Mandarin and Spanish are selected as the target languages. In this study it is assumed that *the unknown input speech data is monolingual*. Three different approaches that perform multilingual phone recognition are investigated and descriptions of these approaches are given in Section 2. Details of the multilingual test systems are given in Section 3. Section 4 presents the phone recognition experiments and results, followed with conclusions in Section 5.

2. MULTILINGUAL PHONE RECOGNITION APPROACHES

For the task of Multilingual Phone Recognition (MPR) the language of the input speech utterance is unknown. Thus, Language Identification (LID) will be performed at some stage during the recognition process in order to produce the final monolingual results. This LID procedure in the system can generally be divided into two different categories: *Explicit-LID* and *Implicit-LID*. Explicit-LID implies that the MPR system performs LID on the unknown input data explicitly. Implicit-LID means that LID is an integral part of the MPR process.

2.1 Explicit-LID – Approach 1

One approach to Explicit-LID employs an external LID system to first identify the language of the input utterance and the corresponding monolingual system is then selected to perform the phone recognition (as shown at Figure 1, Approach 1). This is one of the most straightforward approaches to achieve MPR. When no LID errors occur, this approach achieves the same performance as the monolingual systems. Therefore the accuracy of the external LID system is the main concern to the overall system performance. The advantage of this approach is that it can employ language-dependent speech recognition techniques (e.g. different acoustic and language models) on each monolingual recogniser. However, it can not handle the case where the input utterance contains multiple languages as this system can only give monolingual results.

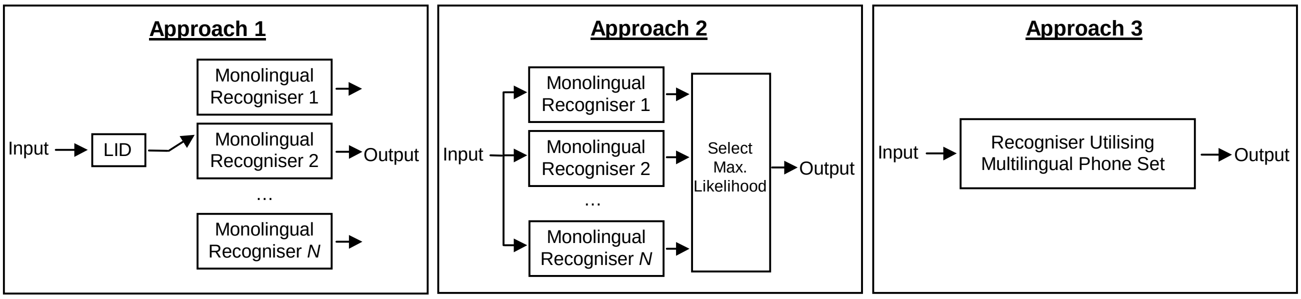

Figure 1. Block diagram of the three different approaches to achieve multilingual phone recognition.

2.2 Explicit-LID – Approach 2

Another approach to Explicit-LID is to run all of the monolingual recognisers in parallel and select the output generated by the recogniser that obtained the highest likelihood score as the final result (as shown at Figure 1, Approach 2). LID in this approach is performed at the end of the MPR process. The selection of final output not only results in the phone transcription, but also identifies the language of the input utterance. This approach shares similar characteristics with Approach 1, differing only in the strategy of the LID process.

2.3 Implicit-LID – Approach 3

A MPR approach that belongs to Implicit-LID utilises a set of multilingual phones within a single recogniser (as shown at Figure 1, Approach 3). These multilingual phones can be created by merging monolingual phones amongst the target languages that share the same phonetic symbol (e.g. IPA [5] and Worldbet [6]) or acoustically similar according to a certain distance measure or similarity criteria.

The aim of this phone merge procedure is to obtain a minimal phone set that covers all of the sounds that exist in the target languages, however, certain language-specific information are removed from the system after the phone merge procedure. The symbols of these merged monolingual phones will be mapped to a new symbol in the transcriptions. Therefore, before process the speech data for training and testing, all phonetic transcriptions will be mapped to the new multilingual phone set. Note that LID is *not* required to perform MPR with this approach, as the generated output can be directly compare to the correct (mapped) transcriptions. Nevertheless, language can be explicitly identified when this approach is applied to multilingual speech recognition. The recognised output in that case will indicate the language of the input utterance at the text level.

The advantage of this MPR approach is the recognition system can be configured to handle input utterances that contain multiple languages, however language-specific speech recognition techniques can not be easily applied as some of the language specific information is merged together (at the phone level) in the system.

3. TEST SYSTEMS

3.1 Baseline Monolingual Phone Recognisers

The OGI-TS database [4] and hidden Markov model toolkit (HTK) [7] are used to train and test the phone recogniser for the languages of English, Mandarin and Spanish. Each feature vector is extracted at every 10ms and comprises Perceptual Linear Predictive (PLP) cepstral coefficients with energy plus it’s delta and acceleration coefficients. Phone model (monophone) topology comprises a 3 state Hidden Markov Model (HMM) with 8 Gaussian mixture components per state.

Original transcriptions, which are transcribed using the Worldbet [6] phonetic symbol set, are modified such that all diacritics are removed. In addition, some phones are mapped to similar sounds to ensure sufficient data coverage. A language tag is appended to each phone symbol to preserve language information while performing multilingual experiments. The amount of training data and the number of phones used in each monolingual recogniser are shown in Table 1.

3.2 Multilingual Phone Recognition Test Systems

The Approach 1 MPR system in this study employs an external LID system that utilises efficient Gaussian Mixture Model (GMM) analysis, as described in [8]. The characteristics of each language is modelled by a single GMM. During LID, the language represented by the GMM that obtained the highest likelihood score against the unknown test utterance is deemed as the identified language. The Universal Background Model (UBM) technique [9] was employed to improve the efficiency of the system during both LID training and testing. In addition, Vocal Tract Length Normalisation (VTLN) was applied to reduce speaker variation presented in the speech data. All GMMs, including the UBM for both VTLN and testing and GMM for each language, are trained using the OGI-TS database. The advantages of this GMM-LID system are that transcriptions are not required for model training and faster than real time performance can easily be obtained on regular computer hardware. The error rate of this GMM-LID system tested on an 11-language experiment with 45-second test segment is around 14% [8].

Language	Training Data (Hour)	# Phone Model	% Correct	% Accuracy	Deletion	Substitution	Insertion	Total Test Phones
Isolated Phone Recognition								
English	1.5	41	49.2	-	-	4103	-	8079
Mandarin	0.5	40	50.5	-	-	2744	-	5539
Spanish	0.9	37	58.0	-	-	3092	-	7357
Continuous Phone Recognition								
English	1.5	41	47.9	32.8	1432	3487	1430	9448
Mandarin	0.5	40	49.6	34.3	1032	2368	1028	6740
Spanish	0.9	37	56.3	45.9	904	2893	904	8689

Table 1. Details and experimental results of the baseline monolingual phone recognisers. Continuous phone recognition results under "% Accuracy" included insertion error while "% Correct" does not.

The Approach 2 MPR test system makes direct use of the baseline monolingual phone recognisers without any modification. It has the simplest structure among the three approaches but has the highest computational cost.

A multilingual phone set is created for the MPR test system that utilised Approach 3 by merging monolingual phones with the same Wordlbet symbol. For each merged phone, a new model is trained with the data pooled from the merged languages. After the phone merge procedure, the number of phones is reduced from a total of 118 to 73. As mentioned at Section 2.3, all phonetic transcriptions are mapped to the multilingual phone set with the original language tag at each phone symbol removed.

4. EXPERIMENTS AND RESULTS

The test data set contains 17 test utterances (continuous speech) for each language and each utterance has a duration of around 45 seconds. Both isolated and continuous phone recognition experiments were performed. For isolated phone recognition, a single phone token is prepared as input for each test system. Each of these tokens are extracted from the test data set according to the transcription, with those tokens with a duration of less than 30ms (or 3 feature vectors) removed. For continuous phone recognition experiments, the entire test utterance is utilised as input to the test systems. A phone-loop recognition network is used to perform the continuous recognition.

Table 1 shows the experimental results of the baseline monolingual phone recognisers along with the number of deletion, substitution and insertion errors for the continuous phone recognition experiment. All results are obtained with approximately the same deletion and insertion rate using the insertion penalty option in HTK. The average number of correctly recognised phones is about 53% for isolated phone recognition and 51% for continuous phone recognition. No direct comparison of the results across the languages can be made as the amounts of train and test data varies. However, it does indicate that phone recognition in Spanish might be an easier task than the other two languages due to lower number of phones.

The experimental results for both the multilingual isolated and continuous phone recognition tasks are shown at Table 2. LID accuracy for Approach 1 and 2 were also included. No LID was required for Approach 3 as mentioned in Section 2.3 and also for the baseline system where the language information is known a priori (or equivalently 100% correct LID).

4.1 Isolated Phone Recognition Results

Input speech data in this task typically contains fewer frames of speech and therefore provided very little information for the system to correctly identify its language. Therefore, multilingual isolated phone recognition is a difficult task and this is supported by the experimental results, where in average 31% of phones are correctly recognised across all the approaches compared to 53% for the baseline system. LID accuracy of Approach 1 and 2 are only 49% and 50% respectively. The phone recognition performance is greatly affected by the poor LID results as these approaches depend on LID to decide the final monolingual output. Approach 3 performs the best amongst all MPR approaches with 36% of phones correctly recognised. This approach does not depend on LID and its decline in performance is mainly contributed by the confusion between the 73 multilingual phone models.

4.2 Continuous Phone Recognition Results

Approximately 45 seconds of speech data is used as input in this experiment. Experimental results indicate that typically a LID accuracy of around 90% was achieved. The exception to this was Spanish, which resulted in 77% for Approach 1 and only 35% for Approach 2. Phone recognition rates for MPR Approach 1, 2 and 3 are 47%, 39% and 36% respectively compared to 51% for the baseline system. The success of Approach 1 hinges on the LID accuracy achieved by the system with an average of 89%. It can be seen that when applied to continuous phone recognition the increase in available input data duration results in an improvement in LID performance and subsequently an improvement in overall phone recognition for Approach 1 and 2. Conversely Approach 3 does not benefit from this increase in utterance duration.

Comparing the experimental results across the languages in Approach 3, the performance drop in Spanish compared to the baseline system was 18%. This compares to the drop of around 11% and 16% for English and Mandarin respectively. One explanation for these results is that the Spanish phone set (37 phones) has a higher degree of confusion when contained within the multilingual phone set (73 phones) compared to English and

MPR Approach	English		Mandarin		Spanish		Average	
	% Correct	% LID	% Correct	% LID	% Correct	% LID	% Correct	% LID
Isolated Phone Recognition								
Approach 1	27.6	(4339/8079) 53.7	27.0	(2639/5539) 47.6	28.3	(3421/7357) 46.5	27.6	49.3
Approach 2	32.8	(4959/8079) 61.4	26.7	(2504/5539) 45.2	29.0	(3258/7357) 44.3	29.5	50.3
Approach 3	38.7	-	32.2	-	36.2	-	35.7	-
Baseline	49.2	-	50.5	-	58.0	-	52.6	-
Continuous Phone Recognition								
Approach 1	43.3	(16/18) 88.9	49.6	(17/17) 100.0	47.0	(13/17) 76.5	46.6	88.5
Approach 2	45.8	(17/18) 94.4	46.4	(16/17) 94.1	24.2	(6/17) 35.3	38.8	74.6
Approach 3	37.3	-	33.5	-	37.9	-	36.2	-
Baseline	47.9	-	49.6	-	56.3	-	51.3	-

Table 2. Multilingual (isolated and continuous) phone recognition experimental results. "% LID" is in percentage of correct.

Mandarin (41 and 40 phones respectively). This is reinforced by the results achieved by Approach 1 and 2 where the LID accuracy for Spanish was similarly poor. Conversely for English and Mandarin, with higher LID accuracy of around 92% and 97% respectively averaged across Approach 1 and 2, phone recognition rate of Approach 3 did not suffer to as large degree. Therefore, although Approach 3 did not perform LID, the language-specific information presented in the data are contributed *implicitly* to the overall performance. In contrast to Approach 1 and 2 where the language-specific information are extracted *explicitly* from the input data, these extracted information plus its correctness contributed mainly to the overall phone recognition performance.

5. CONCLUSIONS

This paper compares three different approaches to achieve multilingual phone recognition and present results from experiments from isolated and continuous phone recognition. Three of the world's most spoken languages; English, Mandarin and Spanish were selected as the target languages. For the isolated phone recognition experiment, MPR Approach 3 (Implicit-LID, which utilises multilingual phone set in a single recogniser) has performed the best with 36% phone recognition rate. It obtains superior results because it does not suffer from the difficulty of identifying the language from a single phone token. In the continuous phone recognition experiment, MPR Approach 1 (Explicit-LID, which utilised an external LID system to select monolingual recogniser) obtaines the best phone recognition rate of 47% compared to 51% for the baseline monolingual system. Its superior performance is mainly due to the higher LID accuracy obtained by the external GMM-LID system in the experiment.

Experimental results indicate that different MPR approaches should be employed for different applications according to the degree of LID accuracy that can be achieved from the input test utterance. If high LID accuracy is achievable, Explicit-LID MPR approach that depends on LID can obtain better performance, otherwise the Implicit-LID MPR approach is more appropriate.

6. ACKNOWLEDGEMENTS

This work is sponsored by a research contract from the Australian Defence Science and Technology Organisation (DSTO). The author would like to thank Terrence Martin, Michael Mason and John Dines for valuable advice provided.

7. REFERENCES

[1] Kohler, J., "Multi-lingual Phoneme Recognition Exploiting Acoustic-Phonetic Similarities of Sounds," *International Conference on Spoken Language Processing*, vol. 4, pp. 2195-2198, 1996.

[2] Schultz, T. and Waibel, A., "Fast Bootstrapping of LVCSR Systems with Multilingual Phonemes sets," *Eurospeech*, vol. 1, pp. 371-374, 1997.

[3] Imperl, B., Kacic, Z., Horvat, B., and Zgank, A., "Agglomerative vs Tree-based Clustering for the Definition of Multilingual set of Triphones," *International Conference on Acoustics, Speech and Signal Processing*, vol. 3, pp. 1273-1276, 2000.

[4] Muthusamy, Y. K., Cole, R. A., and Oshika, B. T., "The OGI multi-language telephone speech corpus," *International Conference on Spoken Language Processing*, vol. 2, pp. 895-898, 1992.

[5] IPA, *Handbook of the International Phonetic Association : a guide to the use of the International Phonetic Alphabet*: Cambridge University Press, 1999.

[6] Hieronymous, J., "ASCII Phonetic Symbols for the World' s Languages: Worldbet," *Journal of the International Phonetic Association*, 1993.

[7] HTK, "The Hidden Markov Model Toolkit (HTK)," http://htk.eng.cam.ac.uk/, 2002.

[8] Wong, E. and Sridharan, S., "Methods to Improve Gaussian Mixture Model Based Language Identification System," *Internaional Conference on Spoken Language Processing*, vol. 1, pp. 93-96, 2002.

[9] Reynolds, D. A., "Comparison of background normalization methods for text-independent speaker verification," *Eurospeech*, vol. 2, pp. 963-966, 1997.