AUTOMATIC LANGUAGE RECOGNITION USING HIGH-ORDER HMMS

J.A. du Preez and D.M. Weber
Department of Electrical and Electronic Engineering, University of Stellenbosch,
Private Bag X1, Matieland 7602, South Africa,
Email: dupreez@firga.sun.ac.za

Abstract

We present automatic language recognition results using high-order hidden Markov models (HMM) and the recently developed ORder rEDucing (ORED) and Fast Incremental Training (FIT) HMM algorithms. We demonstrate the efficiency and accuracy of pseudo-phoneme context and duration modelling mixed-order HMMs as well as fixed order HMMs over conventional approaches. For a two language problem, we show that a third-order FIT trained HMM gives a test set accuracy of 97.4% compared to 89.7% for a conventionally trained third-order HMM. A first-order model achieved 82.1% accuracy on the same problem.

1 Introduction

In a companion paper [1], we developed theory for efficient high-order hidden Markov modelling. In this paper we demonstrate its practical applicability to speech processing. Phonotactic modelling in automatic language recognition (ALR) systems is a large and complex model of the interdependence of the phonemes of each language. Because of the size and complexity of these models, we have chosen this as a suitable field to demonstrate our techniques with. The intention of this paper is to demonstrate the applicability of high-order HMMs trained using ORED/FIT approach to a practical speech system, not to develop a fully-fledged ALR system. In particular, we compare FIT to conventional high-order HMM training and we investigate the effect of specialised context and duration modelling mixed-order HMMs we have developed. We consider the use of ergodic high-order HMMs to model languages as this approach does not require costly phoneme transcription of the speech data. Details of some other ALR systems may be found in [3, 4, 5, 6].

2 Database and signal processing

Our prior work [7] indicated that about 1 hour of speech was adequate for modelling a language with a first-order HMM. Preliminary experiments indicated that the higher-order Markov models would need substantially more data. Since they need not be transcribed, this does not necessarily pose a problem. From the OGI-TS[1] database, roughly 100 minutes of free-format English and Hindi speech was available as training data. These were the two largest collections of data available and were therefore used in the experiments. Silence sections in the recordings were removed automatically by using an energy criterion. The power in the remainder was normalised to compensate for recording volume. After pre-emphasis, tenth-order LPC-cepstra and delta-cepstra features were calculated from 32ms time frames spaced at 16ms intervals. Cepstral mean subtraction was used to compensate for channel variation. For testing data, two independent sets of data were used. These were 5s segments and the 45s NIST'95 LID set. Each language model had its own transition probability description, while one central set of pdfs was referenced collectively by all the language models.

3 HMM structure and training procedure

In an application like language recognition, the left-to-right HMMs commonly used in speech recognition systems are inappropriate as the the HMM can start in any state and jump to any state as dictated by the language model. In this work, we use sixteen state ergodic high-order HMMs, The state transition probabilities are selected for first- and fixed-order HMMs as well as mixed-order models that model both "phoneme" context and duration [1]. Fixed order models of order n are labeled Fn i.e. a second-order fixed model is labeled F2. Context models and duration models are labeled Cn and Dn respectively. We also investigate models that implement both context and duration modelling. These are labeled as CnDm models where n and m refer to the order of the context and duration respectively [2]. The ALR system consists of a bank of such models, each optimised to a specific language. Unknown speech is matched to each of the models and classified according to which one fits best. Prior work [7]

[1] The Oregon Graduate Institute kindly made the OGI-TS database available to us

has shown first-order ergodic HMMs which use a common set of probability density functions (pdf) for all the languages to be modelled significantly enhances robustness. All training was done using the Viterbi re-estimation algorithm [8]. Higher-order models were always reduced to equivalent first-order form by using the ORED algorithm, enabling the use of the first-order re-estimation algorithm in all cases.

The Viterbi algorithm includes a matrix that records the optimal path between states as a function of time. In a first-order HMM system, the product of the number of states with the number of time frames in the speech segment, dictates the size of this matrix. To reduce the demands on memory, the training sequences were subdivided into 5s sequences that formed the basic patterns presented to the system. After training, all transition probabilities leading from the initial state were set to be equal (i.e. 1/16) to ensure that the model could start in any state with equal probability.

4 Comparison between FIT and direct training

We now compare ALR results for fixed-order models trained using the FIT algorithm (train successively higher order models) and direct training (order reduction followed by training). These results will validate results obtained on synthetic data as reported in [1]. FIT trained models are labeled Fn while direct trained models are labeled Xn.

4.1 Computational requirements

The computational costs of the different approaches are measured in terms of the number of transition probabilities, memory requirements) and the number of operations (CPU time) required to process equivalent problems. The results are summarised in Table 1. We find that the computational advantage of the FIT approach becomes significant for third-order models. The final F3 model contained a compact 2301 transition probabilities compared to the 47161 probabilities in the X3 model. These additional transition probabilities translate directly into CPU time and memory costs. From this it is clear that the FIT algorithm results in much more compact models, getting more so as the order increases. This corroborates the results obtained from synthetic data using this approach [1].

Order	MEM	CPU	Size
2	69%	94%	70%
3	13%	7%	5%

Table 1: Comparison of computational requirements and final model sizes for 16-state ergodic HMMs trained via direct and FIT algorithms. The results are expressed as FIT requirements as a percentage of direct requirements.

4.2 Classification accuracy

We now compare the classification accuracy of the FIT and direct training approaches. These results are summarised in Table 2 and Table 3. On the training data, the accuracy of the FIT and direct trained models are comparable. On independent testing data, however, in all cases the FIT trained models perform similarly or better than the direct trained models. Furthermore, all the FIT trained models result in smaller differences between training and testing set accuracies than those achieved by the direct trained models. This, combined with the larger models that resulted from direct training, indicates greater specialisation in such models. With the 5s classification trials, a McNemar test [9] with a 90% significance level shows all the high-order models to be more accurate than the baseline HMM1. Due to the rather small test set, the experiment based on the NIST'95 45s set (39 trials) could only show the 3rd-order FIT model (F3) to be more accurate than the 1st-order HMM on a 90% significance level. No statistically significant differences could be detected between the direct and FIT trained models. The FIT approach thus obtains statistically similar results at a significantly reduced computational cost.

	5s (2169 trials)		45s (339 trials)	
Order	Direct	FIT	Direct	FIT
1	83.4%	-	92.6%	-
2	86.8%	85.2%	95.3%	95.3%
3	92.7%	89.6%	98.8%	99.1%

Table 2: Accuracy measured on training set for 16-state ergodic HMMs trained via direct and FIT algorithms.

	5s (247 trials)		45s (39 trials)	
Order	Direct	FIT	Direct	FIT
1	69.2%	-	82.1%	-
2	75.7%	76.1%	87.2%	89.7%
3	79.8%	79.8%	89.7%	97.4%

Table 3: Accuracy measured on testing set for 16-state ergodic HMMs trained via direct and FIT algorithms.

4.3 ALR using context and duration HMMs

In this section we experiment with fixed-order, and mixed-order (context and duration modelling) HMMs. Figure 1 illustrates the categorisation (context and duration orders) and FIT training dependencies of the different models.

Model 1 is a first-order 16 state ergodic HMM used in the prior experiments. It serves as a base-line model and all

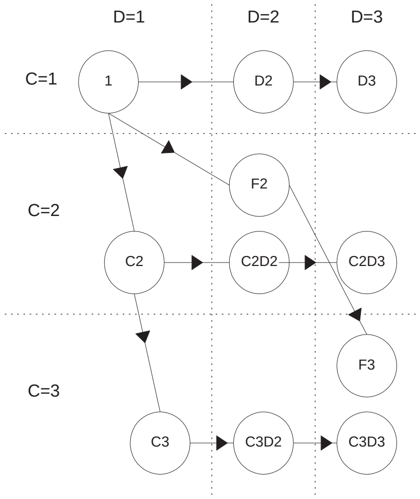

Figure 1: Identities and FIT training dependencies (indicated by arrows) of ALR models. C is the context order and D is the duration order.

other models are extensions of it. The fixed-order model results (F2 and F3) are also directly imported from this previous section. The C2 and C3 models, while neglecting duration modelling ensure that the system "knows" about respectively 2 and 3 distinct prior states when making a transition to a next state. The longer history that is being modelled by maintaining the C values in this way, necessarily makes these models more sensitive to training data deficiencies. Models D1 and D2 add duration modelling capabilities to model 1, while C2D2, C2D3, C3D2 and C3D3 add duration information to the respective models which they extend. DmCn models were not considered since they suffer from reduced context order when trained on repetitive data [2]. We do not suggest that the above topologies are the only viable ones. Alternative topologies are the subject of future research.

Model sizes: We find that the number of transition probabilities in the Cn models are considerably larger that the Fn models investigated. For example, the F3, C3 and C3D3 models have 2301, 4286 and 9070 transition probabilities respectively. This compares favorably to the direct trained X3 model with 47161 transition probabilities. Second-order models (F2, C2, D2 etc) exhibit a similar but less pronounced trend. Each of these different topologies were designed for different purposes. We suspect that the growth in the number of transition probabilities of the Cn models can be attributed to the larger history of states that it is modelling. The longer such a history is, the greater the potential number of state combinations will be. When duration modelling is added to form the CmDn models, it is done on an already enlarged model, thereby contributing to a marked growth in the number of parameters.

Classification results The classification accuracy on the training and test sets are given in Table 4. If we consider the training set accuracy, we note the added benefit of each successive FIT extension. The rapid increase in accuracy of the various context-emphasized models indicates an increasing ability to fit the training data. Tests on an independent set of testing data reveal that the pattern of improvement with each new FIT extension in the training data results is followed. At the context order 2 level (C=2), the context-emphasized models appear to function very competitively. We suspect that our training database was too small to sustain the long history span utilized at the C=3 levels. The large difference in accuracy between the training and testing sets (i.e. specialisation is taking place) for the C3 family also confirms this. In general models benefit from duration modelling. McNemar significance tests reveal almost identical results to those obtained in Section 4.2. In general the high-order models improve on the baseline HMM1, while no statistically significant differences could be detected between the high-order models.

It is safe to conclude that increasing the duration and/or context orders is indeed beneficial, as long the training database is large enough to sustain it. The various CmDn models hold much promise for ALR, but need more testing on larger databases. Although it might seem unfair to compare results from the simple first-order model to that of large high-order models, the purpose of these experiments was to investigate the role of context and duration modelling HMMs, and not to compare models containing an equal number of parameters.

Due to the general lack of standardised ALR databases, direct comparison with prior work is difficult. However, Lund et al. [6] utilize an acoustic-phonetic based scheme that does not require any transcriptions during training and can thus be directly compared to our work. On language pairs from the NIST'95 set, they achieve accuracy ranging from 85.2% to 93.6%. Although based on different principles, the 97% accuracy that we achieved compares well with this. Others using phoneme recognition based systems (thus requiring phoneme transcriptions) have obtained 97.9% [3] and 94.8% [4] on the NIST'94 45s set. The FIT/ORED approach compares well to these results and is particularly attractive as it does not require costly hand transcription of the speech data.

5 Outstanding ALR issues

We view the ALR experiments reported here as as a prototypical demonstration of concept. Many necessary refinements are absent from it and there are also several aspects that require further investigation. In order to expand it into a full-blown ALR system incorporating many languages, at least the following should receive attention:

1. A deeper investigation of various high-order HMM topologies.

2. A thorough investigation of the relationship between

	5s (247 trials)			45s (39 trials; NIST'95)		
Training Set Accuracy						
Ctx\Dur	D=1	D=2	D=3	D=1	D=2	D=3
C=1	1: 83.4%	D2: 84.1%	D3: 86.4%	1: 92.6%	D2: 94.1%	D3: 94.7%
C=2	C2: 89.0%	F2: 85.2% C2D2: 90.3%	C2D3: 91.4%	C2: 98.5%	F2: 95.3% C2D2: 99.1%	C2D3: 99.1%
C=3	C3: 94.6%	C3D2: 95.4%	F3: 89.6% C3D3: 97.3%	C3: 99.7%	C3D2: 100 %	F3: 98.8% C3D3: 100 %
Test Set Accuracy						
Ctx\Dur	D=1	D=2	D=3	D=1	D=2	D=3
C=1	1: 69.2%	D2: 79.4%	D3: 80.6%	1: 82.1%	D2: 92.3%	D3: 92.3%
C=2	C2: 75.7%	F2: 76.1% C2D2: 78.1%	C2D3: 77.7%	C2: 92.3%	F2: 89.7% C2D2: 92.3%	C2D3: 94.9%
C=3	C3: 76.9%	C3D2: 76.1%	F3: 79.8% C3D3: 76.5%	C3: 89.7%	C3D2: 89.7%	F3: 97.4% C3D3: 94.9%

Table 4: Training and Testing set classification accuracy for different orders and types of HMM models.

the size of the models and the size of a database sufficiently large to train it is necessary.

3. Interpolation techniques are commonly used in N-grams applications [3] to mitigate poor training set size should be investigated.

4. The incorporation of gender has shown to increase accuracy and should thus be investigated.

5. Previous work [5] indicates the usefulness of explicitly removing the acoustic component from the resultant score when matching unknown speech to a language model. Although our system uses a common acoustical model for the languages concerned, removing it altogether from the final matching score should be investigated.

6 Conclusion

This paper shows that the FIT algorithm is indeed practical for training large real-life high-order HMMs. It confirms the benefits over direct training that we found in previous simulation experiments. The FIT algorithm provides greater computational efficiency, results in more compact models and if anything, increases accuracy. We demonstrated some very promising prospects for implementing ALR systems that do not need transcriptions. We also demonstrated techniques that achieve independent control over context modelling (modelling sequences of consecutive states while ignoring their individual repetitions) and duration modelling (modelling the duration/repetitions for which a specific state is active) in high-order HMMs.

References

[1] Du Preez, J.A. and Weber, DM. “Efficient high-order hidden Markov modelling”, *Proceedings of the IEEE International Conference on Speech and Language Processing*, 1998.

[2] Du Preez, J.A., *Efficient high-order hidden Markov modelling. PhD Dissertation*, University of Stellenbosch, South Africa, 1998. URL: http://dsp.ee.sun.ac.za/reports

[3] Zissman, M.A. “Language identification using phoneme recognition and phonotactic language modelling”. *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 3503 - 3506. Detroit, USA, 1995.

[4] Kadambe, S. and Hieronymus, J.L. Language identification with phonological and lexical models. Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, pp. 3507 - 3510. Detroit, USA, 1995.

[5] Mendoza, S., Gillick, L., Ito, Y., Lowe, S. and Newman, M. “Automatic language identification using large vocabulary continuous speech recognition”. *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 785 - 788. Atlanta, USA, 1996.

[6] Lund, M.A., Ma, K. and Gish, H. “Statistical language identification based on untranscribed training”. *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 793 - 796. Atlanta, USA, 1996.

[7] Du Preez, J.A. “Language Recognition by means of Ergodic Hidden Markov Models”. *Proceedings of the IEEE COMSIG Conference*, pp. 33 - 38, Cape Town, South Africa, 1992.

[8] Levinson, S.E. “Structural methods in automatic speech recognition”. *Proceedings of the IEEE*, vol. 73 no. 1, pp. 1626 - 1650, 1985.

[9] Gillick, L. and Cox, S.J. “Some statistical issues in the comparison of speech recognition algorithms”. *Proceedings of the IEEE International Conference on Acoustics Speech and Signal Processing*, pp. 532 - 535, Glasgow, UK, 1989.