Proceedings of the Institute of Acoustics

VARIABLE POOL SIZE, TIED PARAMETER SYSTEMS FOR CONTEXT-DEPENDENT SUB-WORD UNIT SPEECH RECOGNITION.

David Ollason

BT Labs, Martlesham Heath, Ipswich, Suffolk, IP5 7RE, UK

1. INTRODUCTION

In automatic speech recognition the transition from a context-independent to a context-dependent modelling system almost inevitably results in an increase in the number of models to be trained. As a consequence of this, the amount of training data afforded to each model can become grossly inadequate. This trade off between model specificity and model trainability has been the subject of much research for a number of years [2,3,4]. One trend in this research has been to develop context-dependent HMM sub-word unit systems which compensate for the lack of allophone training data by tying the parameters of each group of allophones, associated with the same monophone, to a separate pool [5,6]. Systems which keep the number of shared model parameters constant across groups of tied allophones fail to address two important aspects of the training problem. Firstly there is typically a large variance in the distribution of the amount of training data across the allophone groups and secondly the variability of sounds to be modelled within each group is different.

The approach presented in our paper extracts information from the training data, prior to model generation, and uses it to select the number of parameters shared by each group of models.

The remainder of this paper is organised as follows. In Section 2 the method used to select model topology is outlined. The contruction of the context-independent models used as seed models for the tied parameter systems is described in Section 3. Section 4 details the algorithms used to generate the three types of tied parameter system investigated. The algorithm used to synthesise triphone models from biphone models is presented in Section 5. Both the evaluation database and signal parameterisation are described in Section 6. Following this, Section 7 outlines the experiments performed to evaluate these systems. Finally a discussion of the relative merits of each system and some conclusions are included in Section 8.

2. MODEL TOPOLOGY SELECTION

Our approach aims to link the number of parameters shared by each group of models to the trainability of that model group and also to the variability within the group. As a measure of trainability we use the number of training tokens available for the allophone group and we define variability as the number of different triphone contexts present within the group. In this set of experiments we have chosen to vary the parameter pool size according to the Equation 1.

$$S_m = F \times \sqrt{T_m} \times V_m \qquad \text{Equation 1}$$

S - Parameter pool size for allophone group m.

F - Constant which is used to control the total number of parameters.

T - Number of training tokens associated with allophone group m.

V - Number of triphone contexts present in group m (measure of variability within group).

This equation has been applied to three different types of tied parameter system. The task used to evaluate the performance of this technique was the speaker-independent recognition of examples of 697 different surnames collected over the UK telephone network.

3. CONTEXT INDEPENDENT SEED MODELS

Sets of context-independent models were required as seed models for the tied parameter systems and also to provide baseline context-independent performance levels. The context independent symbol set used for this work comprised 44 vocabulary symbols and one noise symbol. Two main systems were created based on this symbol set. The first was a 3 state 12 mode system containing 1584 modes in total. The second system was a 3 state variable mode system also with a total of 1584 modes. This system was created by applying Equation 1 during the allocation of modes to each state of each context-independent unit. In this case each context-independent unit is treated as an allophone group. The performance of these systems is presented in Table 1, Section 7.1.

4. TIED PARAMETER SYSTEMS

The following sections describe the three types of tied parameter system investigated in this work.

4.1 Clustered State System (CSS)

In this system the corresponding states of all the allophones derived from the same monophone are shared in common pools. The application of Equation 1 to this system results in a variation in the size of the shared state pool from one allophone group to the next. Every state in the system has a fixed number of modes. This arrangement is similar to the one presented in [5] although the methods employed to control the number of parameters in each cluster are different.

Figure 1 shows the arrangement for a single allophone group with 1 mode per state

where the number of corresponding states has been reduced from 3 to 2 by the state clustering process.

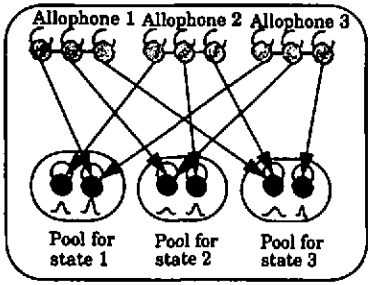

Figure 1: Clustered State System

4.2 Shared Mode System (SMS)

In this system the corresponding states of all the allophones, derived from the same monophone, share a common pool of modes. The application of Equation 1 to this system results in a variation in the size of the shared mode pool from one allophone group to the next.

Figure 2 shows the arrangement for a single allophone group where the corresponding states in each allophone share a pool of three modes. Allophones within a group differ only in their mode weights.

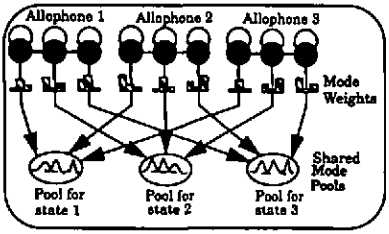

Figure 2: Shared Mode System

It is interesting to compare this system to a context-independent system. The allophones in the shared mode system are simply separate sets of mode weights, acting on a common pool of modes. In terms of the shared mode system the context-independent system can be viewed as a single allophone with one set of mode weights acting on the same modes.

4.3 Shared Mode & Clustered State System (SM&CSS)

This system can be thought of as a combination of the two described previously. All the allophones derived from the same monophone share a common pool of modes and a common pool of states for corresponding states.

Figure 3 shows the arrangement for a single allophone group where the number of corresponding states has been reduced from 3 to 2 by the state clustering process, and the states in each pool share a pool of three modes. Comparing Figure 3 to Figure 2 one can see that the two systems are essentially the same except that the shared mode and clustered state system has fewer sets of mode weights. State clustering applied to a shared mode system effectively results in the clustering and sharing of sets of mode weights whilst the modes themselves remain unaltered.

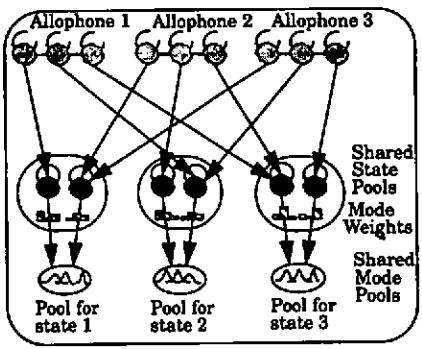

Figure 3: Shared Mode & Clustered State System

5. TRIPHONE CONSTRUCTION

The recognition network for this task contains 2210 different triphones. Of these 23% occur 3 or fewer times in the training corpus and only 14% have 50 or more occurrences. For this reason it was decided that for each experiment two shared parameter systems would be constructed, one for left biphones and one for right biphones. Following the re-estimation of both shared parameter systems the triphones required for recognition are synthesised from the two biphone systems [1].

6. DATABASE

The training corpus used for these experiments consisted of 13448 examples of isolated spoken surnames collected over the UK telephone network from several hundred callers spread throughout the country. The test data was a separate set comprising 494 examples of spoken surnames collected in the same manner from a different set of talkers. The utterances in the training set were automatically seg-

mented using phoneme models trained on the Subscriber database [8]. There were in total 697 different surnames in the recognition network. Some of the surnames had multiple pronounciation transcriptions resulting in a network size of 1000. The 95% confidence interval for this task is ± 4.3% based on an assumed error rate of 58%.

7. EXPERIMENTS

In order to be able to compare each of the shared parameter systems the total number of modes in each system was kept constant and equal to the number of modes in the context-independent systems described in Section 3. Although the number of modes is kept constant the number of mode-weights, and therefore the total number of parameters, varies from system to system.

All the recognition times quoted are relative to the recognition time for the standard fixed mode context-independent system shown in row 1 of Table 1. The timing information was acquired by applying the UNIX command *time* to each run of the HTK recognition tool *HVite*.

All vocabulary models are 3 state left to right with no skips. A single state 6 mode noise model was used throughout.

7.1 Context Independent Experiments

For comparison with the context-dependent systems Table 1 shows the results achieved by a 3 state 12 mode context-independent system (CI-FIX) and also those achieved by a 3 state variable mode context-independent system (CI-VAR) containing the same total number of modes.

Table 1: Context-Independent,results

Model Type	Percentage Accuracy	Recognition Time
CI-FIX	54.7	1.0
CI-VAR	56.3	0.8

7.2 CSS Experiments

A set of single mode and a set of four mode, context-independent models were generated. Using these context-independent models as seed models two separate context-dependent, clustered state model sets were created.

Using the single mode context-independent models as seed models a separate CSS was generated for left and right biphones. The final number of clustered states is set such that the total number of modes in the system is 1584 (CSS-1M).

Using the four mode context-independent models as seed models a separate CSS was generated for left and right biphones. The final number of clustered states was set to one quarter that in the CSS-1M such that again the total number of modes in the system is 1584 (CSS-4M).

The CSSs generated were then given one iteration of embedded re-estimation. The triphones required for recognition were synthesised from the two biphone CSSs.

The results for sets 1&2 are shown in Table 2

Table 2: CSS results

Model Set	Percentage Accuracy	Recognition Time
CSS-1M	58.5	1.2
CSS-4M	53.4	1.3

7.3. SMS Experiments

A set of variable mode and a set of fixed mode, context-independent models containing 1584 modes in total were generated. Using these context-independent models as seed models two separate context-dependent, shared mode model sets were created.

Using the fixed mode context-independent models as seed models a separate SMS was generated for left and right biphones. This results in a SMS where each pool of shared modes is the same size, in this case 12 (SMS-FIX).

Using the variable mode context-independent models as seed models a separate SMS was generated for left and right biphones. This results in a SMS where the pools of shared modes are of different sizes (SMS-VAR).

The SMSs generated were then given one iteration of embedded re-estimation. The triphones required for recognition were synthesised from the two biphone SMSs.

The results for sets 1&2 are shown in Table 3.

Table 3: SMS results

Model Set	Percentage Accuracy	Recognition Time
SMS-FIX	58.5	2.8
SMS-VAR	60.9	3.2

7.4 SM&CSS Experiments

A set of variable mode, context-independent models containing 1584 modes in total were generated. Using these context-independent models as seed models three separate context-dependent, shared mode & clustered state model sets were created for left and right biphones.

The number of states in the first system was clustered to 3/4 the original number (SM&CSS3/4), in the second system to 1/2 (SM&CSS1/2) and in the third to 1/4 SM&CSS1/4).

The SM&CSSs generated were then given one iteration of embedded re-estimation. The triphones required for recognition were synthesised from the two biphone SM&CSSs.

The results for systems 1,2&3 are shown in Table 4.

Table 4: SM&CSS results

Set	Percentage Accuracy	Recognition Time
SM&CSS3/4	61.3	1.8
SM&CSS1/2	59.7	1.7
SM&CSS1/4	58.5	1.5

8. CONCLUSIONS

There are clear advantages in terms of accuracy to gained from modelling contextual variation within sounds. Further gains in accuracy are made when this technique is coupled with a modelling structure which reflects the amount of training data available and the variability within the sounds to be modelled. This work also demonstrates that it is possible to produce significant accuracy improvements without the dramatic computational overhead normally incurred with context-dependent modelling.

Compared against standard fixed mode, context-independent modelling the three shared parameter systems investigated produced differing performance in terms of both recognition accuracy and computational efficiency.

Clustered state systems produced a gain of 3.8% (from 54.7% to 58.5%) with a corresponding rise by a factor of 1.2 in recognition processing time. This accuracy increase may not be statistically significant.

Shared mode systems were found to give significant gains in terms of recognition accuracy at the cost of increased computation at recognition time. Recogni-

tion accuracy rises by 6.2% (from 54.7% to 60.9%) with a corresponding rise of by a factor of 3.2 in recognition processing time.

Shared mode, clustered state systems succeed in dramatically reducing the computational overhead of the shared mode systems whilst maintaining the accuracy gains achieved. Recognition accuracy rises by 6.6% (from 54.7% to 61.3%) with a corresponding rise by a factor of 1.8 in recognition processing time.

Taking both recognition accuracy and processing time into consideration it would appear that the Shared Mode, Clustered State system provides the best overall performance.

A lack of training data and model inflexibility are two of the limiting factors in determining how well the contextual variations within a phoneme can be modelled. This work has shown that variable pool size, shared parameter systems provide increased model flexibility, allowing the models to make more efficient use of the limited training data available.

9. REFERENCES

[1] Young S.J.,*"HTK Version 1.3: Reference Manual"*, Cambridge University Engineering Dept, Speech Group, May 1992.

[2] Hwang M., Huang X. D., Alleva F., *"Predicting Unseen Triphones With Senones"*, Proc ICASSP April 1993, pp 311-314, Minneapolis.

[3] Hwang X. D., Jack M. A., *"Semi-Continuous Hidden Markov Models for Speech Recognition"*, Comp. Sp. and Lang., 3, pp 239-251, 1989.

[4] Lee K. F. *"Large Vocabulary Speaker Independent Continuous Speech Recognition: The SPHINX system"*, Ph.D Thesis, CMU-CS-88-148, 1988.

[5] Young S.J., *"Benchmark DARPA RM Results With The HTK Portable HMM Toolkit"*, Proc DARPA Workshop, September 1992, Stanford.

[6] Paul D.B., *"The Lincoln Continuous Speech Recognition System: Recent Developments And Results"* DARPA Speech & Natural Language Workshop, February 1989, pp. 160-166, Philadelphia.

[7] Wood L.C. and Pearce D.J.B., *"Sub-Word HMM Recognition: An Investigation Of Phone Context Modelling And Improved Discrimination"*, Proc. IOA Conf. on Speech and Hearing, Windermere, pp. 181-188, 1990.

[8] Simons A.D., Edwards K., *"Subscriber - A Phonetically Annotated Telephony Database"*, Proc. IOA Conf. on Speech and Hearing, Windermere, pp. 9-16, 1992.