MODELLING PHONETIC CONTEXT USING HEAD-BODY-TAIL MODELS FOR CONNECTED DIGIT RECOGNITION

Janienke Sturm, Eric Sanders

A²RT, Dept. Language and Speech, University of Nijmegen, The Netherlands
{sturm,eric}@lands.let.kun.nl

ABSTRACT

Both whole word modelling and context modelling have proven to improve recognition performance for connected digit strings. In this paper we will show that word boundary variation can be effectively modelled by applying the Head-Body-Tail (HBT) method as proposed by Chou et al in [1] and also applied by Gandhi in [2]. Each digit is split into three parts, representing the beginning, middle and end of a word. The middle part - the body - is assumed to be context-independent, whereas the first part - the head - and the last part - the tail - incorporate information about the preceding or subsequent digit. The results we obtained with HBT-modelling are compared with results obtained with whole-word models (WWM's) [3] and with the results obtained with HBT-models reported in [2]. It is shown that using HBT models a relative improvement over context-independent WWM's of 28% on string level can be reached.

1. INTRODUCTION

Pronunciation variation is an important issue in speech recognition research. Pronunciation does not only depend largely on the speaker, but also (within speakers) on the direct context of the word. Since automatic speech recognisers suffer from this variability, a lot of research on recognition of spoken language has aimed at reducing the effect of these types of variability on the performance of the speech recogniser. Many researchers have successfully attempted to model pronunciation variation by training context-dependent phonemes (i.e. triphone models) and by adding pronunciation variants to the lexicon, the language model and/or the train corpus [4]. In connected digit recognition whole word models (WWM's) are often used. Whole word modelling in this domain is feasible, because of the limited size of the lexicon. Using whole word models a lot of within word variation is already accounted for in the acoustic models. However, a lot of pronunciation variation occurs at word boundaries: the first and final sounds of words are to a large extent dependent on the following and preceding words. In order to deal with variation at word boundaries, context dependency has to be introduced to the acoustic models. Training context-dependent whole word models would require an enormous amount of training data, due to the large number of models that would have to be trained: for each word model at least 10 left-hand contexts and 10 right-hand contexts need to be modelled, which adds up to 1,000 word models. In order to make context modelling for digit recognition feasible, we made use of the Head-Body-Tail (HBT) method which was proposed by Chou et al in 1994 [1], and successfully applied by Gandhi in 1998 [2]. In this approach each digit is split up into three parts. The middle part of the word - the body - is assumed to be context-independent, whereas the first part - the head - and the last part - the tail - are dependent of the previous and subsequent digit respectively. Thus, for each digit 10 heads, 10 tails and one body are trained. This way, the total number of models that needs to be trained is reduced from 1000 to 210.

In order to create the best performing HBT models for our digit recognition task we experimented with different model topologies and we varied the number of Gaussian densities that were trained for each model. In this paper we will present the results of these experiments. We will compare the results with results of context-independent whole-word models [3]. The results will also be compared with the HBT results obtained by Gandhi et al [2].

The paper is further organised as follows. Section 2 describes the speech material used for training and evaluating the models, the topology of the models and the language model. It also describes the experiments that we conducted. In Section 3 the results are presented, which will be discussed in Section 4. Finally, in Section 5 we will draw conclusions and discuss future research.

2. METHOD AND MATERIAL

2.1. Train and test material

The speech material that was used to train and evaluate the acoustic models consists of connected digit strings, spoken in Dutch, selected from three databases:

a. SESP: a speaker verification database consisting of connected digit strings, ranging in length from 1 to 16 digits, with an average length of 7.3. This database consists of pin-codes (4 digits) telephone numbers (10 digits) and scope card numbers (14 digits). All utterances were recorded over the telephone network in various environments.
b. POLYPHONE digits: this part of the POLYPHONE database consists of strictly connected digit strings, ranging in length from 1 to 16 digits, with an average length of 6. The database was recorded over the fixed telephone network.
c. CASIMIR: this database consists of pin-codes (four digits) and scope card numbers (14 digits), spoken strictly as connected digits. The average length is 7.7.

All data were collected over the telephone network in various environments.

For training 9,753 utterances were selected from the SESP and POLYPHONE databases. The training set was composed carefully according to the following criteria:

- the Signal-to-Noise ratio (SNR) is larger than 10 dB,
- the clipping rate is equal to 0,
- number of digits per string is not equal to 1 (too short), 10 or 14[1],

Furthermore, the training set was balanced according to the number of occurrences per digit, the number of occurrences at the start or end of a string and the number of digits and utterances per sex.

The test set consists of 10,000 utterances selected randomly from the CASIMIR database and from the remaining utterances of SESP and POLYPHONE.

2.2. Acoustic models

For each digit one context-independent body model is trained, together with 11 context-dependent head models and 11 context-dependent tail models: one for each preceding or subsequent digit and one for preceding or subsequent filled pauses or silence exceeding 250 ms.. The latter contexts are modelled because we do not expect context dependency to extend over filled pauses and silences of this length. In addition to the digit models, one noise model consisting of three states was trained for different kinds of background noise and filled pauses. Since our recogniser has a built-in silence-detector, we did not explicitly train a silence model.

In total 231 acoustic models were trained. Each model consists of at least three states (the total number of states varied in different experiments). The total number of digit variants in the lexicon is 1,211.

2.3. Feature extraction

Feature extraction of all 8 kHz-sampled speech files was done using a 16 ms. Hamming window with a 10 ms. shift. Of each speech sample 14 Mel Frequency Cepstral Coefficients (MFCCs) and their first order derivatives were calculated, i.e.28 features.

2.4. Language model

Since our recogniser does not support the use of a grammar, we used a probabilistic bi-gram language model to enforce the correct combinations of heads and tails. It is important to note that, with our recogniser, using a probabilistic language model means that false combinations of heads and tails can only be made improbable, not impossible.

The language model was created by summing all possible combinations of heads and tails and assigning probability scores to these combinations. In our experiments the distribution of bi-gram probability scores reflects the distribution of occurrences in the train corpus.

2.5. Experiments

A number of experiments were carried out to test the performance of different HBT models.

First, a base-line experiment (EXP1) was carried out in which all heads, bodies and tails consist of three states, and, consequently, each digit of nine states. In EXP2 and EXP3, the total number of states per digit is duration dependent. The number of states in these experiments is based on the mean, minimum and maximum duration of the digit as observed in the train corpus. The total number of states per digit in these experiments ranges from 10 (/e:n/ 'one') to 19 (/ne:x@n/ 'nine'). In EXP2 only the body is duration-dependent, whereas the heads and tails for all digits consist of three states. In EXP3 also the heads and tails are duration-dependent, here the total number of states is distributed evenly over the head, body and tail models. All models were trained with both 32 and 64 Gaussian densities per state. Due to the large number of models, especially for the heads and the tails, 64 Gaussian models are probably undertrained due to shortage of training data. Therefore, in EXP4 the number of Gaussian densities per state for the bodies is 64, whereas for the heads and tails we experimented with 8 (EXP4.1), 16 (EXP4.2) and 32 (EXP4.3) densities per state. The number of states in these experiments is the same as in EXP2. The results of all experiments are shown in the next section.

3. RESULTS

This section presents the results of the experiments described in section 2.5. All results are presented in terms of Word Error Rate (WER) and String Error Rate (SER), where:

- WER = (# insertions + deletions + substitutions / total # words) * 100%
- SER = (# strings containing one or more errors / total # strings) * 100%.

In order to determine the optimal number of Gaussian densities trained per model, the baseline experiment (EXP1) in which all models consist of three states, was carried out using 4, 8, 16, 32 and 64 Gaussian models. Table 1 shows the results for EXP1.

# Gauss	Total # Gauss	WER	SER
4	2,761	12.34%	48.90%
8	5,481	8.44%	37.02%
16	10,839	7.51%	33.03%
32	21,356	7.31%	32.44%
64	40,930	6.10%	27.73%

Table 1 Performance of the base-line models as a function of the number Gaussian densities per model

The data in Table 1 show that error rates drop significantly when more Gaussian densities are trained.

[1] Both telephone numbers and scope card numbers have a very fixed format.

The second column shows, however, that not for all models the maximum number of Gaussian densities has been trained. Table 1 also shows that the performance of the base-line models does not seem very good for the CDR task. This is probably due to the fact that an equal number of states is used for all models.

In EXP2 and EXP3 the number of states for the acoustic models depends on the duration of the digits. In EXP2 only the body is duration-dependent, in EXP3 this also applies to the heads and the tails. Table 2 shows the results for 32 and 64 Gaussian models. For ease of comparison, the first two rows in Table 2 show the performance of the base-line models.

Experiment	# Gauss	WER	SER
EXP1	32	7.31%	32.44%
	64	6.10%	27.73%
EXP2	32	3.52%	16.53%
	64	3.42%	16.08%
EXP3	32	3.84%	18.05%
	64	3.82%	18.02%

Table 2 Performance of duration-dependent models in EXP2 and EXP3 for 32 and 64 Gaussians

As can be seen in Table 2, the models that take duration into account perform significantly better than the base-line models. Also, assigning three states to all heads and tails (EXP2) works better than distributing the states evenly over head, body and tail (EXP3). In these experiments the difference between the 32 Gaussian and 64 Gaussian models is not significant, which indicates that with this amount of training data the maximum number of Gaussians that can effectively be trained is 32.

Table 3 shows the results for EXP4. In this experiment, the number of Gaussians that was trained for each head and tail was reduced to 32 (EXP4.3), 16 (EXP 4.2) and 8 (EXP4.1), whereas the total number of Gaussians for the body models remained 64. The model topology used in this experiment is the same as in EXP2, the performance of these models (64 Gaussians) is repeated in the first row of Table 3.

Experiment	Total # Gauss	WER	SER
EXP2	42,790	3.42%	16.08%
EXP4.1	9,589	3.25%	15.71%
EXP4.2	14,812	3.23%	15.62%
EXP4.3	25,162	3.22%	15.60%

Table 3 Performance of models with different combinations of Gaussians

Table 3 shows an improvement for EXP4.1, EXP4.2 and EXP4.3, compared to EXP2. This confirms our assumption that the 64 Gaussian head and tail models were undertrained due to shortage of training data. The results also show that not much is gained from moving from 8 to 16 or 32 splits for the heads and tails. This seems to indicate that with the size of our train set, the maximum resolution of the models is already reached with 8 Gaussians per state.

To put the results into perspective, Table 4 shows the results obtained with context-independent WWM's. These models were trained on the train set described in section 2.1, using the same total number of states as in EXP2. Table 4 also shows the results of both WWM and HBT models obtained by Gandhi.

Experiment	WWM	HBT	% Improvement
A^2RT	17.23%	15.60%	9.4%
Gandhi	12.12%	9.04%	25.4%

Table 4 Performance of our models and the CI and HBT models of Gandhi (SER)

Table 4 shows that although our HBT models perform better than the context independent whole word models, the improvement is much smaller than the improvement achieved by Gandhi. Possible explanations will be discussed in the next section.

4. DISCUSSION

Table 1 in the previous section shows that the performance of the HBT models increases when models are used with more Gaussian densities per state. The reason for this is that those models are able to more adequately model the variation. However, training this many densities requires a lot of training data. From the figures in Table 1 we could conclude that there seems to be enough training material to train as much as 64 Gaussians. However, it may very well be that the gain in performance over the 32 Gaussian models is almost only due to the body models. Since there are only 10 body models, the total number of training samples per models is approximately 6,000, whereas for the head and tail models only 300 occurrences are available, which might be close to the minimum. This idea is supported by the data in Table 3: combinations of less Gaussians for the head/tails and more for the body result in a small improvement of 3% on string level. Furthermore, moving from 8 to 16 or 32 Gaussians for the heads and tails does not improve the recognition results significantly, which also indicates that there is not enough training material to perform extra splits on the head/tail models.

Comparing the performance of the HBT models with the performance of the context-independent WWM's, we see a relative improvement of 9.4%. Compared with the improvements Gandhi observed in his experiments (25.4%), our gain in performance thanks to the HBT models is rather small. Also, the absolute error rates are much higher. There are a number of possible explanations for this observation.

First, a very important difference between Gandhi's experiment and ours is that since our recogniser does not support the use of a grammar an n-gram language model

was used to enforce the correct combinations of heads and tails. As mentioned earlier, in our recogniser, by using a probabilistic language model instead of a restrictive grammar, combinations can only be made less likely, but not impossible. This introduces unnecessary errors: in EXP4.3 almost 500 of all recognised pairs of heads and tails were incorrect. Incorrect in this case means that either the tail does not match the next digit, or the head does not match the previous digit. In order to investigate the influence of the grammar on the performance of the HBT models, we repeated EXP4.3 (which gave the best results) using a recogniser that does support the use of a grammar. In this grammar we explicitly enforce the correct combinations of heads and tails. Since this is a strict grammar, no other combinations can occur in the recognition result. The grammar does not contain any information about the number of digits per string. Feature extraction, training procedure and model topology were kept similar to our previous experiments. Results (SER) of this experiment, using 32 Gaussians for heads and tails en 64 for the bodies, are shown in Table 5.

# Gauss	LM	Grammar	% Improvement
EXP4.3	15.60%	12.33%	21.0%

Table 5 Performance of recognisers using a language model and using a grammar

Table 5 shows that using a grammar a relative reduction of the SER of 21.0% can be achieved. Given the fact that all parameters were kept as similar as much as possible, this reduction can almost completely be ascribed to the use of a grammar. The total relative improvement over the context independent WWM's using the grammar is 28.4%, which is comparable to the improvement achieved by Gandhi.

Second, there are a number of differences between the acoustic model sets. As mentioned earlier, in section 2.2, the model set that we trained consists of 221 head, body and tail models. Given the size of our train set, and considering the results in Table 3, this might be not enough to adequately train this amount of models with 16 or more Gaussians. The obvious solution would be to add more data to the train corpus. Another solution, which has also been applied by Gandhi, is to apply tying. Since a number of heads and tails are very much alike and behave the same in different contexts, these can be joined in one model. This reduces the total number of models to be trained, and therefore increases the amount of training data available per model. Another difference is the way noise is modelled. In Gandhi's experiments, two models are used to absorb noise, one representing breath and mouth noises and one more general filler model. In our method only one model was used for all kinds of filled pauses and background noise. In our experiments, using the best models, 49% of all insertions are due to noises being mistaken as a digit.

Finally, another difference is the way the acoustic models were trained. In our experiments the acoustic models are trained using Maximum Likelihood training (ML). In order to use the training material more efficiently, Gandhi applied Minimum Classification Error training (MCE). MCE reduces the number of mis-classifications made during training by applying a special cost function This results in more 'correct' training data for each model. Gandhi showed that, for the WWM's, using MCE the performance of the models increases with 17.4% on string level. Since MCE training has also been applied to the HBT models, they are likely to perform better.

Further improvement of the HBT models can be achieved in a number of ways.

In [3] it is shown that a relative reduction of 20% in SER can be achieved by training separate models for male and female speech. Since this result was achieved using the same total number of Gaussian densities trained, the reduction of the number of training data per model is not a problem.

Another way to improve the models can be found in the training procedure. All models in our experiments have been trained on the basis of a linear segmentation of the train material, using a silence-speech detector and taking into account the number of states of each model. In order to improve the segmentation of the speech material, and therewith the accuracy of the models, it may be better to start the training from a bootstrap segmentation obtained using the best performing HBT models.

5. CONCLUSIONS AND FUTURE WORK

In this paper we have shown that, for recognition of Dutch digit strings, context can effectively be modelled using the Head-Body-Tail approach. Results indicate that using a combination of 32 Gaussian models for the heads and tails and 64 Gaussian models for the bodies results in a string error rate reduction of 9.4% over context independent WWM's. Using a grammar to restrict the number of possible combinations of heads and tails further reduces the error rate with 21.0%. This way a total relative improvement of 28.4% is achieved.

Future research will be aimed at improving the performance of the HBT models. A number of ways to achieve this were proposed in the discussion.

6. REFERENCES

1. Chou, W., Lee, C.-H., Juang, B.-H., "Minimum Error Rate Training of Inter-Word Context-Dependent Acoustic Model Units in Speech Recognition", *Proceedings ICSLP'94*, 1994.
2. Gandhi, M.B., Jacob, J., "Natural Number Recognition using MCE Trained Inter-Word Context-Dependent Acoustic Models", *Proceedings ICASSP'98*, 1998.
3. Scharenborg, O., Bouwman, G., Boves, L., "Connected Digit Recognition with Class Specific Word Models", *Proceedings VOTS2000*, 2000.
4. Strik, H. Cucchiarini, C., "Modeling pronunciation variation for ASR: a survey of the literature", *Speech Communication,* Vol. 29:225-246,1999.