Radboud Repository

Radboud University Nijmegen

PDF hosted at the Radboud Repository of the Radboud University Nijmegen

The following full text is an author's version which may differ from the publisher's version.

For additional information about this publication click this link.
http://hdl.handle.net/2066/75043

Please be advised that this information was generated on 2019-02-15 and may be subject to change.

CONNECTED DIGIT RECOGNITION WITH CLASS SPECIFIC WORD MODELS

Odette Scharenborg, Gies Bouwman, Lou Boves

A^{2}RT, Dept. Language & Speech, University of Nijmegen
P.O. Box 9103, 6500 HD Nijmegen, The Netherlands
{odettes, bouwman, boves}@lands.let.kun.nl

ABSTRACT

This work focuses on efficient use of the training material by selecting the optimal set of model topologies. We do this by training multiple word models of each word class, based on a subclassification according to a priori knowledge of the training material. We will examine classification criteria with respect to duration of the word, gender of the speaker, position of the word in the utterance, pauses in the vicinity of the word, and combinations of these.

Comparative experiments were carried out on a corpus consisting of Dutch spoken connected digit strings and isolated digits, which are recorded in a wide variety of acoustic conditions. The results show, that classification based on gender of the speaker, position of the digit in the string, pauses in the vicinity of the training tokens, and models based on a combination of these criteria perform significantly better than the set with single models per digit.

keywords: connected digit recognition, acoustic modelling, language modelling

1. INTRODUCTION

Speaker-independent connected digit recognition (CDR) over the telephone is a particularly interesting challenge for automatic speech recognition. On the one hand, the size of the vocabulary is small, which should make the task tractable. On the other hand, a digit string is incorrect when only one digit is recognised incorrectly. Therefore, string lengths of ten or more require a per digit recognition accuracy close to 100% in order to keep the string recognition accuracy higher than, say, 98%. Optimal use of the available training material and training techniques are of crucial importance to reach this 'near perfect' recognition accuracy.

The focus of the work presented here is efficient use of the training material by selecting the optimal set of models and their topologies. Efficient use of the material means finding the number of models, states and densities that maximises performance. It is known that training just one model per phone or word is not always optimal. Many digit recognisers use separate model sets for male and female speakers. In addition, the authors in [1] proposed to train models for fast, average, and slow realisations of the words. In [2],[3] realisation speed and speaker gender were combined in order to train gender dependent word models, for fast and slow realisations of the training tokens separately. In all cases, significant recognition improvements were reported.

These studies suggest that prior knowledge of the training material can be used to improve recognition performance. In [4] it was shown that a Classification Tree approach to the problem proves that linguistic features can be used to advantage. In this paper we investigate whether comparable improvements can be obtained with a rule based or 'common sense' approach. In doing so, we investigate two features (viz. the position of a digit in a string and the presence of a pause before or after a digit) that have not been used before for the purpose. In summary, we will examine classification criteria with respect to

- duration of the digit,
- gender of the speaker,
- position of the digit in the string,
- pauses in the vicinity of the digit, and
- combinations of these.

Different criteria will result in different numbers of models per digit, different numbers of states, and eventually different numbers of Gaussian densities. In order to allow a fair comparison we will keep the total number of densities in all model sets roughly equal. A system with just 10 models, but with a high number of densities per state will serve as the reference.

This paper is organised as follows. In Section 2, we explain the different selection criteria on which the class specific models are based. Section 3 presents the results of the experiments. In Section 4, we give an interpretation of these results. Finally, in Section 5 we summarise our method, briefly draw the most remarkable conclusions and outline some of our plans for follow-up research.

2. METHOD

We measure the influence of each classification by comparing the performance of a speech recognition system using class specific models to a baseline system with only 10 models. In the remainder of this paper we will refer to this model set as BASE. All model sets investigated in this paper represent whole word models. All models have the same left-to-right HMM topology, but the number of states for each model is one of the optimisation parameters.

The general procedure for training class specific word models is as follows:

1. add a label to each word in the baseline transcription of the training corpus according to the subclass imposed by the current classification criterion;
2. determine the duration distribution of each subclass in order to choose the number of states for each model, using a forced alignment with the BASE models;
3. generate a uni- and bigram language model based on the labels in the transcription.

First, we explain the five selection criteria in more detail. Section 2.6 and 2.7 then elaborate on the second and third step.

2.1 Digit duration

To account for different speaking rates, between speakers and within speakers, we trained duration based models.

The median of the duration distribution of the digit was taken as threshold value to divide the set of digit tokens into short and long realisations, thus, both sets have an equal amount of training tokens. To this aim the following labels were added to the digit tokens in the transcription:

short for digit tokens comprising fewer frames than the median number of frames of that digit type and
long for digit tokens comprising at least as many frames as the median number of frames of that digit type.

We will use shorthand notation DUR to refer to this model set.

2.2 Speaker gender

The training databases used in this study contain only utterances labelled for speaker gender. This allows us to add gender labels to the words in the transcription:

male for words uttered by male speakers and
female for words uttered by female speakers

This model set will be referred to as GENDER.

2.3 Word position

Many phonetic and ASR studies have shown that the acoustic realisation of words is strongly affected by the position of the word in an utterance. For example, string final digits tend to have a falling pitch contour, lower intensity and longer duration. This motivates a distinction between three subclasses per digit, indicated by the following labels:

initial for the first digit in an utterance,
middle for digits from the second up to the penultimate digit, and
final for the last digit in an utterance.

A consequence of this definition is that in case the average string length of the training corpus is greater than three, the *middle* set contains more tokens than the *initial* and *final* sets. Single digit utterances obtain the *final* label, because their acoustic properties resemble those of final digits most. In the remainder of this paper this set will be denoted as POS.

2.4 Pause context

The final criterion for distinguishing between models is the presence of a pause in the vicinity of the digit. Most speakers tend to cluster long digit strings into groups of two, three or four digits, separated by short pauses. It is not unlikely that this clustering of strings into small groups affects the acoustics and duration as well, as already pointed out in [5]. Therefore, each digit is given one of three labels:

head for a digit preceded, but not followed by a pause,
between for a digit neither preceded, nor followed by a pause, and
tail for any digit followed by a pause

In our experiments we consider a pause as a silence of at least 250 ms. Each utterance is considered to be preceded and followed by a pause. Digits surrounded by pauses are labelled with a *tail* tag, for the same reason why we labelled POS for isolated digits as final. We will use PAUSE as shorthand notation for this model set.

2.5 Combination of criteria

In addition to the criteria presented in the previous paragraphs, it is also possible to combine two or three of them. The order in which to apply the criteria may become important if the criteria are somehow correlated. We examined the following combinations:

- Classification with respect to digit duration, followed by classification for speaker gender. (notation: DUR-GEN)
- Classification with respect to speaker gender, followed by digit duration. (notation: GEN-DUR)
- A combined classification of speaker gender and presence of pauses in the digit context. (notation: GEN-PAUSE)

The first two combinations are examined to investigate whether there is a correlation between the speaker gender and the digit duration. In [2] and [3] the second combination has been investigated for Italian and English digit strings. The difference between the two combinations lies in the number of states defined for each word model. The last combination of criteria was chosen because GENDER models and PAUSE models ranked among the best criteria tested.

2.6 Model topology

Choosing an appropriate number of states for a word HMM is especially important for the experiments with the DUR models. On the one hand, a model with too small a number of states is not capable of modelling the dynamic acoustics accurately, because too many frames are allocated to the same state. On the other hand, models with a number of states much larger than the observed number of frames in the shortest tokens may result in poor modelling during training, because some frames in the vicinity of these tokens will be assigned to the head and/or tail states of these models.

The number of HMM states was set equal to the minimum observed duration, i.e. number of frames, of each subclass in the training material. The duration was determined by a forced alignment of signal and transcription, using model set BASE. The number of states of these baseline models was determined on the basis of a forced alignment with the best phone models available at the start of the research.

2.7 Language Model

For the experiments described in Section 3 a combined uni- and bigram language model was used. The language models were trained on the corresponding transcriptions of the training corpus.

The classification strategies for acoustic modelling, as proposed in the previous subsections, do not necessarily benefit equally from N-gram language models. For the POS models it is unlikely that the bigram language model will add much value. After all, the assumed distinction is purely of an acoustic nature and the language model may put too much restriction on the choice of the best acoustic model. However, for the GENDER models the bigram language model can be expected to add the extra knowledge that during one utterance the models of only one gender must be used. The different contributions of the language model make it an interesting topic to explore. Therefore, we performed tests with and without the language model.

3. RESULTS

Experiments were carried out on a corpus created from three Dutch spoken connected digit databases: Polyphone, SESP and Casimir. All these corpora contain telephone speech recorded in a wide variety of acoustic conditions. The acoustic features were 14 Mel-scale Frequency Cepstrum Coefficients (c0 …c13), and their first order derivatives, i.e. 28 features. These vectors were based on 16 ms frames and a 10 ms frame shift. Next, HMMs were trained. Each state comprised a mixture of maximally 128 Gaussian densities. The training set consisted of 9753 utterances with an average of 6.3 digits per utterance. The unseen test corpus contained 76,682 digits in 10,000 digit strings. Additional information can be found in [6].

The distribution of training material of each criterion is displayed in Table 1.

Model set	Percentage training tokens per subclass
DUR	*short*: 50%, *long*: 50%
GENDER	*male:* 53%, *female*: 47%
POS	*initial:* 16%, *middle*: 68%, *final*: 16%
PAUSE	*head*: 28%, *between*: 35%, *tail*: 37%

Table 1 Distribution of the training tokens for each subclass per model set.

Table 2 shows the word and sentence error rates obtained in the tests we performed with the system with just one word model per digit class (BASE) for 32, 64 and 128 Gaussians per state.

Tot. Gaussians	WER (%)	SER (%)
3744 (5 splits)	4.65	21.78
7481 (6 splits)	4.36	20.56
14920 (7 splits)	4.17	19.63

Table 2 The performance of the BASE models at word and sentence level as a function of the total number of Gaussians per set of models.

Table 3 displays the word and sentence error rates obtained in the tests we performed with the class specific models. For ease of reference the performance of the BASE models is repeated.

Criterion	Tot. Gaussians	WER (%)	SER (%)
BASE	14920	4.17	19.63
DUR	28316	4.20	19.95
GENDER	18877	3.27	15.59
POS	29100	4.52	20.81
PAUSE	29818	3.37	16.54

Table 3 The performance of the class specific models (max. 64 Gauss. / state) as a function of the type of classification criterion.

Table 4 presents the performance of the class specific models, without any kind of language modelling. Again, for ease of reference, the performance of the BASE models is shown in the 2nd row of this table.

Criterion	WER (%)	SER (%)
BASE	4.17	19.63
GENDER	3.36	16.23
POS	3.41	16.30
PAUSE	3.13	14.97

Table 4 The performance of the class specific models (64 Gauss. / state) without a language model as a function of the type of classification criterion.

As can be seen in Table 3 and 4 the performance of GENDER deteriorated in the tests without a language model, while the performance of both POS and PAUSE improves significantly (at a 95% confidence level).

Table 5 displays the word and sentence error rates obtained in the tests with the class specific models for combined criteria, with a language model. There are six models per digit in PAUSE-GEN. Although the individual model sets PAUSE and GENDER have the lowest error rates (cf. Tables 3 and 4), the performance of the combination is much worse.

Criterion	Tot. Gaussians	WER (%)	SER (%)
BASE	14920	4.17	19.63
DUR-GEN	15171	3.33	16.18
GEN-DUR	15664	3.32	15.75
PAUSE-GEN	15495	4.08	20.38

Table 5 The performance of the class specific models (max. 16 Gauss. / state) with a language model as a function of the type of combined classification criterion.

Finally, Figure 1 shows all Word Error Rates as a function of the total number of densities per model set. The dotted line connects the results of the BASE models with 32, 64 and 128 Gaussians per state.

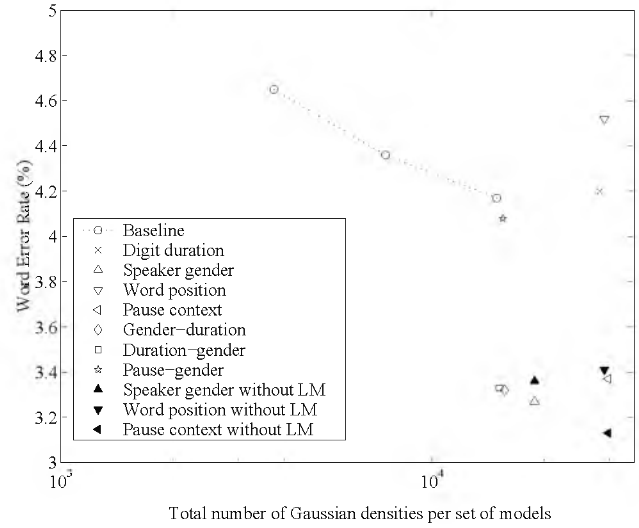

Figure 1 Word Error Rate plotted as a function of the number of Gaussian densities for all tested model sets.

4. DISCUSSION

A fair comparison of the word and sentence error rates can only be made, if the acoustic resolution of the complete set of models is taken into account. This capacity depends on the number of acoustic parameters that have been trained. Therefore, the most efficient model set is the set that uses as few parameters as possible to get a lowest possible error rate.

Keeping this in mind the class specific model sets can be compared to the set of BASE models in Figure 1. Although we did not test systems with single models per digit for exactly the same number of acoustic parameters as the class specific models, extrapolating the BASE performance suggests that it won't drop far below 4.0% WER for a higher number of acoustic parameters.

The results show that all class specific models, except model set DUR, provide better acoustic modelling compared to BASE models. It is remarkable to see that the PAUSE models perform equally well as the well-known GENDER models. However, the performance of these three model sets is strongly dependent on the relative contribution of the language model, as we already predicted in Section 2.8. The word error rates for the model sets POS and PAUSE drop significantly when the language model influence is reduced. These results suggest that the language model may have been too restrictive.

Remarkable is that the performance of our model set DUR is far below the performance of the duration based models in [1,2,3]. One explanation could be that our algorithm to define the number of states for each subclass model is sub-optimal. This is subject for further study.

Concerning the combined selection criteria GEN-DUR and DUR-GEN, the small difference in the number of Gaussian densities are caused by the order in which the selection criteria were applied. This can be explained by the fact that the median of the number of frames for digits spoken by male and female speakers is not always the same. Since our model topology algorithm takes the minimum duration in ms divided by 10 as the number of HMM states, this will result in different model topologies for long duration digit models for male and female speech. However, the error rates are still very much alike, indicating that the order for classification does not matter significantly.

The results obtained with the model set PAUSE-GEN show a clear deterioration in comparison with the individual model sets PAUSE and GENDER. In order to understand this deterioration, we performed an analysis on an independent development corpus. It appeared that the overlap between the set of incorrectly recognised words of PAUSE and that of GENDER is very high. Therefore, it is less likely that combining the classification criteria of PAUSE and GENDER would add much value to either one of the individual model sets. On the other hand, the intention to keep the total number of densities approximately fixed resulted in models with only 16 densities per state. This may not be enough to properly represent all variation within the subclasses.

5. CONCLUSIONS

We compared several classification criteria to select a set of model topologies to make efficient use of the available training material. The classification criteria were word duration, gender of the speaker, word position in the string, and presence of pauses in the vicinity of the digit.

One of the best experimental results presented in this work was obtained with the well-known gender classification criterion. The proposed criterion, for pauses in the vicinity of the training tokens, performed equally well. All class specific model sets, except for the one based on duration, give significant efficiency improvement when compared to the set with single models per digit.

Currently we are experimenting with new ways of defining the number of states per subclass model. The first results are very promising.

6. REFERENCES

[1] Pfau T., Ruske G., "Creating Hidden Markov Models for Fast Speech", *Proc. of ICSLP '98*, Sydney, paper 255, pp. 205-208

[2] Chesta C., Laface P., Ravera F., "HMM Topology Selection for Accurate Acoustic and Duration Modelling", *Proc. of ICSLP '98,* Sydney, vol 7., pp. 2951-2954

[3] Chesta C., Laface P., Ravera F., "Connected Digit Recognition Using Short and Long Duration Models", *Proc. of ICASSP '99,* Phoenix, vol 3., pp. 775-778

[4] Reichl W., Chou W., "Decision Tree State Tying based on Segmental Clustering for Acoustic Modeling", *Proc. of ICASSP '98,* Seattle, vol. 2, pp. 801-804

[5] Godfrey J., Ganapathiraju A., Ramalingam C., Picone J., "Microsegment-Based Connected Digit Recognition", *Proc. of ICASSP '97,* Munich, vol. 3, pp. 1755-1758

[6] http://lands.let.kun.nl/A2RT/cdr/webref.html