CONNECTED DIGIT RECOGNITION USING INDUCTIVE INFERENCE

A. Samouelian

Speech Technology Research Laboratory
Department of Electrical and Computer Engineering
University of Wollongong

ABSTRACT-This paper proposes a novel approach to connected digit recognition by the use of inductive inference "decision trees". To develop the production rules, the expert is bypassed and instead the classification models are generated inductively by examining a large speech database and then generalising the pattern from the specific examples. This approach has already been successfully used for isolated digit recognition [Samouelian, 1996]. The aim of this research is to demonstrate that inductive learning can provide a viable alternative approach to existing automatic speech recognition (ASR) techniques. The proposed system uses mel frequency cepstral coefficients (MFCC) front-end signal processing technique . The C4.5 inductive system [Quinlan, 1993] generates the decision tree automatically from labelled examples in the training database. The recognition is performed at the frame level, using an inference engine to execute the decision tree and classify the firing of the rules. A sorting routine is then used to identify the digit string. Connected digit recognition results for Texas Instruements (TI) digit database, for speaker dependent and independent recognition, for known and unknown digit string lengths are presented.

INTRODUCTION

Connected digit recognition plays an important part in many applications such as telephone number inquiry, credit card validation, personal identification number (PIN) entry, database enquiries, to name a few. Current connected digit recognition systems (Bush and Kopec, 1987; Lee and Rabiner, 1989; Bhurke *et al.*, 1994) use sophisticated algorithms for recognising spoken digit strings using whole word models. These algorithms are generally based on maximum likelihood recognisers such as Hidden Markov Modelling (HMM) and Neural Networks (NN).

This paper proposes a novel approach to connected digit recognition by the use of inductive inference "decision trees". To develop the production rules, the expert is bypassed and instead the classification models are generated inductively, that is by examining a large speech database and then generalising the pattern from the specific examples. This approach has already been successfully used for isolated digit recognition [Samouelian, 1996].

The proposed ASR (Samouelian, 1994a, 1994b) system uses mel frequency cepstral coefficients (MFCC) front-end signal processing technique. The C4.5 inductive system [Quinlan, 1993] generates the decision tree automatically from labelled examples in the training database. The classification is performed at the frame level, using an inference engine to execute the decision tree and classify the firing of the rules. A sorting routine is then used to identify the digit string.

The aim of this research is to demonstrate that inductive learning can provide a viable alternative approach to existing automatic speech recognition (ASR) techniques. The proposed approach has the ability to generate decision trees using any combination of features (parametric or acoustic-phonetic). This allows the integration of features from existing signal processing techniques, that are currently used in HMM stochastic modelling, and acoustic-phonetic features, which have been the cornerstone of traditional knowledge based techniques. Connected digit recognition results for Texas Instruements (TI) digit database, for speaker dependent and independent recognition, for known and unknown digit string lengths are presented.

TRAINING AND RECOGNITION STRATEGY

Speech database

The speech database used for training and testing consisted of the adult speakers from the connected digit subset of the large isolated and connected digit database developed by Texas

Instruements [Leonard & Doddington, 1984]. The database was dialectically balanced American English and was collected in a quiet environment and sampled at 20 kHz. The connected digit subset contained 55 utterances by each speaker, consisting of two repetitions of each of the 11 digit strings each of length 2, 3, 4, 5 and 7. There were 225 speakers, 111 men and 114 women. The database was divided into training and testing subsets as shown in Table 1.

Speaker No.	Sex	Classification
55	Male	Train
56	Male	Test
57	Female	Train
57	Female	Test

Table 1. Training and testing subsets of the database

To avoid the time consuming task of segmenting the training database by hand, a HMM connected digit recognition system developed for TI connected digit database was used to automatically segment and generate appropriate label files at the word level. The small number of misrecognised speech files were removed from the database. These label files were used to automatically label the training database.

For the development and evaluation of the system, the following subsets of corpus TI were used :

- TI-TRNM-28(a): The first repetition subset of TI-TRN(ab) using the training utterances of the first 28 male speakers (1536 training utterances by 28 speakers).
- TI-TRNF-28(a): The first repetition subset of TI-TRN(ab) using the training utterances of the first 28 female speakers (1539 training utterances by 28 speakers).
- TI-TRNM-TST-28(a): The first repetition subset of TI-TRNM(ab) (1485 training utterances from the remaining 27 male speakers).
- TI-TRNF-TST-28(a): The first repetition subset of TI-TRNF(ab) (1595 training utterances from the remaining 29 female speakers).
- TI-TSTM-28(a): The first repetition subset of TI-TSTM(ab) (3080 test utterances from the 56 male speakers).
- TI-TSTF-28(a): The first repetition subset of TI-TSTF(ab) (3135 test utterances from the 57 female speakers).

The connected digit system contained two decision trees, trained separately on male and female subsets of the training data. During the testing of the system, the gender of each speaker was assumed to be known and the appropriate decision tree was used for classifying the utterance.

Front-end signal processing

The speech was downsampled to 8 kHz, pre-emphasised by a first-order filter ($1\text{-}0.95z^{-1}$) (Song and Samouelian, 1993). The signal was then processed by a 256 point (32 ms) Hamming window with a frame shift of 100 points (12.5 ms). A 256 point FFT was performed on each windowed speech portion. A set of 19 mel scaled triangular filter bank was applied to the FFT power spectrum. The log-energy of the 19 outputs was calculated and transformed to 12 mel frequency cepstral coefficients (MFCC) by discrete cosine transform (Davis & Mermelstein, 1980). In order to enhance the distortion measure used in the recognition system, raised cosine function was used to weight the acoustic features. Twelve delta MFCCs were also computed to represent the transitional information. Finally, a delta energy term (time derivatives of the normalised energy) was included taking the total number of coefficients to 25.

Training

A block schematic of the training and recognition strategy is shown in Figure 2. During the training phase, the feature extraction framework extracted the MFCC parameters from the speech signal on a frame by frame basis. The time aligned word labelled files for each digit string were then used to associate each frame with its corresponding label and generate a training data file. The data file contained labelled examples in the form (X,a), where X is the feature vector and a is the corresponding class. This data file was then used by the C4.5 program to generate a decision tree. Only connected digit strings were used for training and testing.

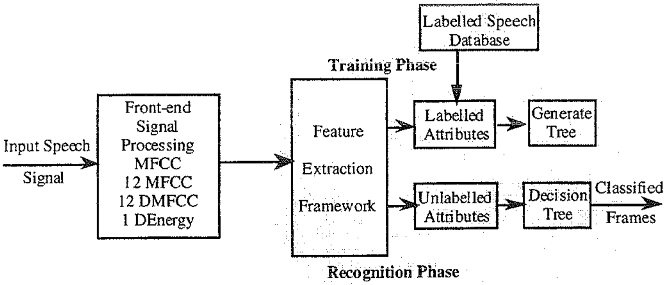

Figure 1. Block schematic of training and recognition strategy

Recognition

Ideally, to recognise a connected digit string, we need to first segment the digit string into individual digits, and then recognise these individual digits. In practice this is not a trivial task since the word boundaries are often fuzzy . We also generally do not know the number of digits in the string, although we may know the range e.g. one to seven digits strings.

In practice, the way we solve the connected digit recognition problem is to find the optimum sequence of word reference patterns to match a given input spectral sequence of vectors through the use of some form of level building algorithm or procedure (Rabiner and Juang, 1993). In recognition systems where a template or reference for each digit is generated from the training database, the dynamic time warping algorithm is used to compare the test utterance against each reference. In stochastic recognition systems such as HMM, the log probability associated with each digit is used in the Viterbi search algorithm to build up the possible connected digit strings. Furthermore, in HMM, generally several states are used to represent each digit.

In inductive learning, a single decision tree is generated to represent all the digits. Thus each digit is represented by a single state and there is a single reference model to represent all the digits. The unknown utterance is classified at the frame level and there are no associated log probability values to help in the search and level building stages. The main problem that needs to be solved is how to go from a frame level classification to a word level classification without the use of dynamic time warping procedure to align the reference with the unknown utterance, and how to determine the possible digit string without the use of Viterbi search . Obviously this is not a trivial problem.

In an effort to solve this problem, the following approach was utilised:

1. Correct minor errors in the unknown utterance at the frame level.
2. Examine for digit clusters to identify possible digit boundaries.
3. Use duration information obtained from the training data to check the digit boundaries, perform digit splitting (e.g. 22, 44 etc.) if necessary and build the possible digit string.

The method used to eliminate simple errors at the frame errors is shown in Table 2.

Errors	Digit sequence	Corrected digit sequence
single	AXA	AAA
two consecutive	AAXYAA	AAAAAA
double	AAXXAA	AAAAAA

Table 2 shows the error correction at the frame level

In addition, to constrain further the possible number of digits in the digit string of the unknown utterance, two cases were allowed for:

1. UL: Unknown digit string length within a nominated range, e.g. one to seven digit for telephone numbers.
2. KL: Known digit length at priori.

RECOGNITION RESULTS

The recognition system was evaluated on 3 subsets of corpus TI. These were the training database, the rest of the speakers in the training data that were not used for training (Test1) and the TI test data (Test2). Note that only utterances "a" were used for training and testing. All of the results are for the combined male and female decision trees. Two cases were evaluated:

1. Unknown digit string length from one to seven.
2. Known digit length of 2, 3, 4, 5 and 7.

The performance of the recognition system was evaluated using HResults program supplied in HTK Toolkit. This allows the correct alignment of the recognised digit string with its corresponding label files and provides word correct, word accuracy, digit string accuracy, digit insertions and deletions. These are calculated as follows:

$$N = H + S + D \qquad \%\text{correct} = \frac{H * 100}{N} \qquad \%\text{Accuracy} = \frac{(H - I) * 100}{N}$$

where:

N = Total number of digits
H = Number of correct digits
S = Number of digit substitutions
D = Number of digit deletions
I = Number of digit insertions

Table 3 shows a summary of the recognition results for unknown digit length (UL) for digit strings from 1 to 7 digits, on the training data (closed-test), while tables 4 and 5 show a summary of the recognition results on Test1 and Test2 data (open-test).

Data	% Correct	% Accuracy	Ins.	Del.	Subs.	%String Accuracy
man	92.8	85.9	444	315	153	52.1
woman	89.5	82.1	478	549	127	49.5
All	91.2	84.0	922	864	279	50.8

Table 3 Summary of recognition results for UL on the training data

Data	% Correct	% Accuracy	Ins.	Del.	Subs.	%String Accuracy
man	82.7	79.8	179	746	323	52.1
woman	76.8	75.8	187	943	475	49.5
All	79.8	77.8	366	168	798	50.8

Table 4 Summary of recognition results for UL on Test1 data

Data	% Correct	% Accuracy	Ins.	Del.	Subs.	%String Accuracy
man	84.3	77.1	932	1091	923	39.4
woman	80.3	73.5	898	1367	1213	34.1
All	82.3	75.3	1830	2458	2136	36.7

Table 5 Summary of recognition results for UL on the Test2 data

Table 6 shows a summary of the recognition results for known digit length (KL) of 2, 3, 5 and 7 digit strings on the test data (Test2) (open-test).

Data	KL2	KL3	KL4	KL5	KL7
% Correct	82.0	79.7	80.8	80.6	80.2
% Accurate	81.3	77.8	77.9	77.1	74.9
%String Accuracy	67.4	50.4	41.9	34.8	17.9

Table 6. Summary of recognition results for KL on Test2 data

Tables 7 show the combined digit confusion matrix for TI-TST data (Test2).

	1	2	3	4	5	6	7	8	9	zero	oh	Del	% Cor.
1	2100	11	5	16	4	6	14	11	20	9	6	112	95.4
2	22	1855	14	10	14	40	26	22	18	24	20	200	89.8
3	7	16	2133	9	5	8	11	26	5	4	6	81	95.7
4	27	6	9	2146	6	6	3	7	7	9	23	53	95.4
5	24	12	7	21	1920	16	21	18	88	19	22	113	88.6
6	12	20	16	16	14	1916	50	39	12	17	18	55	90.0
7	35	21	17	24	20	24	1806	19	44	23	19	137	85.0
8	9	15	41	14	16	20	22	1909	20	11	13	197	91.2
9	29	20	12	29	78	17	34	8	1757	14	17	213	87.1
zero	10	10	12	4	4	18	14	11	6	2154	26	56	95.0
oh	28	19	17	38	32	4	46	15	36	834	1660	230	85.4
Ins	112	124	65	91	92	184	138	95	171	80	165		
% Acc.	85.9	76.4	89.5	89.3	80.1	75.7	73.7	79.8	71.1	89.2	68.8		

Table 7. Confusion matrix for speaker independent recognition on TI-TST(a) data

PERFORMANCE EVALUATION

The results of the extensive system performance evaluations using corpus TI indicate that the proposed system still needs further development to achieve the digit and string accuracy that can be obtained from HMM and NN systems (e.g. better than 98% digit accuracy and 96% digit string accuracy).

Table 7 shows that the recognition performance for UL was relatively uniform across most digits except for the digits *five*, *seven*, *nine* and *oh* , with an average %correct digit recognition of 86.5%, while the remaining seven digits had an average %correct digit recognition of 92.8%. The overall average %correct digit recognition was 82.3% .

Table 5 shows that for UL the overall string accuracy was 36.7% on the test data (Test2), while table 6 shows that for KL, the string accuracy progressively degrades from KL2 to KL7 digit strings. This is mainly due to the increased number of deletions and insertions. It is interesting to note that the %correct digit recognition is relatively constant across the different digit strings for KL strings.

DISCUSSION

The performance of the proposed system falls a long way short of current stochastic techniques and thus there is no need to compare with these approaches. However the performance results indicate that it is possible to use inductive learning for connected digit recognition.

In the proposed approach, a single decision tree is generated to represent all the digits. Each digit is represented by a single state, since there is no automatic clustering procedure that will allow the use of several states per digit. However, it is possible to segment each digit automatically into several states if we use some form of acoustic-phonetic information to subdivide the digit. This should improve the frame level classification by better discrimination between similar sound elements between digits.

The %correct digit recognition rate should improve further by using the whole of the training data instead of a subset of 28 speakers. The major improvement that needs to be made is in the digit string accuracy. This is dependent on the number of insertions and deletions that in turn are dependent on the duration information. There is a need to develop further the digit clustering technique so that number of insertions and deletions are minimised.

CONCLUSION

This paper demonstrated a novel inductive learning, connected digit recognition system using the MFCC front-end signal processing technique. The experimental results indicate that inductive learning can generate decision trees using parametric features instead of the traditional acoustic-phonetic feature set. It can also classify connected digit strings at the frame level. The challenge ahead is to develop a suitable dynamic time warping and Viterbi search algorithm that can provide a path from frame level classification to word level determination and finally build the probable digit string.

REFERENCES

Bhurke, E. R., Cardin, R., Normandin, Y., Rahim, M. and Wilpon, J. (1994). *Application of vector quantised hidden markov modelling to telephone network based connected digit recognition,* Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing, Adelaide Australia, April 1994, pp I105-I108.

Bush, M. A. and Kopec, G. E. (1987). *Network-based connected digit recognition,* IEEE Trans. on Acoustic, Speech and Signal Proc., Vol. ASSP-35, No, 10, pp. 1401-1413.

Davis S. and Mermelstein P. (1980). *Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences,* IEEE Trans. Acoust., Speech, Signal Processing, Vol. ASSP-28, No. 4, pp. 357-366, Aug. 1980.

Lee, C. H. and Rabiner, L. R. (1989). *A frame synchronous network search algorithm for connected word recognition,* IEEE Trans. on Acoustic, Speech and Signal Proc., Vol. ASSP-37, No, 11, pp. 1649-1658.

Leonard R. G. and Doddington G. (1984). *A database for speaker-independent digit recognition, Proc.* IEEE Int. Conf. Acoust., Speech, Signal Processing, San Diego, CA, Mar. 1984, Vol. 3, p. 42.11.

Quinlan, J. R. (1993). *C4.5 programs for machine learning,* Morgan Kaufmann series in machine learning, Morgan Kaufmann publishers, USA, 1993.

Rabiner, L. & Juang, B. H. (1993). *Fundamentals of speech recognition,* Englewood Cliffs., NJ: Prentic-Hall.

Samouelian, A. (1994a). *Knowledge based approach to consonant recognition,* Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing, Adelaide Australia, April 1994, pp I77-I80.

Samouelian A. (1994b). *Knowledge based approach to speech recognition,* Fifth Australian Int. Conf. on Speech, science and Technology, Perth, Australia, pp 479-484, Dec 6-8., 1994.

Samouelian A. (1996). *Isolated voiced digit recognition using inductive inference,* to be published in Proc. IEEE Region Ten Conference, to be held in Perth, Australia, Nov 27-29, 1996.

Song J. M and A. Samouelian (1993). *A Robust Speaker-Independent Isolated Word HMM Recognizer for Operation Over the Telephone network,* Speech Communication, Vol 13, 1993, pp 287-295.