HAL
archives-ouvertes.fr

On The Use of High Order Derivatives for High Performance Alphabet Recognition

Joseph Di Martino

▶ To cite this version:

Joseph Di Martino. On The Use of High Order Derivatives for High Performance Alphabet Recognition. International Conference on Acoustics Speech and Signal Processing - ICASSP 2002, 2002, Orlando, Florida, United States. 4 p, 2002. <inria-00099412>

HAL Id: inria-00099412
https://hal.inria.fr/inria-00099412

Submitted on 18 Jun 2013

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

ON THE USE OF HIGH ORDER DERIVATIVES FOR HIGH PERFORMANCE ALPHABET RECOGNITION

Joseph di Martino

LORIA, B.P 239 Vandœuvre-lès-Nancy 54506 France
E-mail: jdm@loria.fr

ABSTRACT

In this paper I propose new feature vectors for automatic speech recognition. They are based on Mel-cepstrum vectors augmented by derivatives. In the literature, many systems using just two derivatives —delta and delta delta— are described. But none explores the use of higher order derivatives. This paper presents alphabet recognition results on the Isolet database, using feature vectors containing up to the fifth-order derivatives. For this paper I did not use the HTK toolkit proposed by Cambridge University. I developed my own HMM system. I show that with vectors incorporating all the derivatives up to the fifth one, **97.54%** mean recognition accuracy was achieved, result which is comparable to the best published one on this database (**97.6%**), if the recognition accuracy confidence interval concerning this task (approximately 0.3%) is taken into account. It is important to note that this result was obtained without segmenting the speech files by an endpoint detection algorithm. This is an unfavourable experimental condition compared to previous published research works. As a consequence, my system is one of the most powerful systems ever implemented for alphabet recognition.

1. INTRODUCTION

Developing high performance algorithms for alphabet or digit recognition is an important topic nowadays. There are numerous applications of alphabet-digit recognition as for example, spelled name and address recognition, telephone number recognition, managing through the telephone banking accounts etc. Numerous systems have been described in the literature concerning the Isolet database. For example [2] proposed a 2-stage, phoneme-based, context dependent HMM system and 97.37% of recognition accuracy was achieved. Recently [4] and [5] presented new features based on temporal cepstral trajectories. The best mean result was 97.6%. This was obtained with vectors of 50 features. The authors reported that the database was segmented by means of an endpoint detection algorithm that increased their results by 0.3%. Conversely, such a technique was not used in my experiments.

The main contribution of this paper is to show that by using the very popular delta computation paradigm, very good results can also be achieved on a difficult database. This point is not astonishing, because a lot of information concerning the context is extracted from the speech signal with derivatives. A similar scheme was analysed by Zahorian et al. using block feature sub-vectors [3].

While Karnjanadecha's system [5] necessitates less computation and less memory than mine, (50 features instead of 72 for my best results), I do not have in my system the complications generated by the ad hoc Karnjanadecha block expansion strategy. I shall come back on these points in the discussion of the pros and cons of the two systems.

My system is based on an original HMM recogniser that I developed in my laboratory (Loria). The system is described in the next section. This paper is organised as follows. In section 3, I explain how the different derivatives are calculated and consequently how the feature vectors were organised. In section 4, experimental results are presented. In section 5, a discussion on the advantages and disadvantages of my system compared to Karnjanadecha's is given. Finally, concluding and perspective remarks are drawn in section 6.

2. DESCRIPTION OF THE HMM RECOGNISER

The system is based on a classical HMM representation for the letters. Each letter is described by a 5 states, 4 Gaussian mixtures, full covariance matrix, left to right HMM model. Figure 1 illustrates the type of HMM model used.

In the training phase, every training utterance of a letter was segmented in 5 equal parts. All vectors corresponding to a given part were gathered. Then, these vectors were categorised in 4 classes by means of a k-means algorithm. For each class, a covariance matrix and a mean vector were calculated.

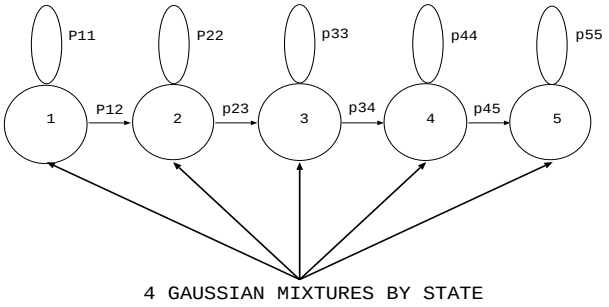

Fig. 1. The HMM model used.

The state transition probabilities were evaluated by equation 3.

Let be X be a random variable giving the mean number of times a HMM state is visited, and assume that the probability distribution of X is given by:

$$P(X = n) = p^{n-1} * q \tag{1}$$

where p is the probability to stay in the same state and $q = 1 - p$ is the probability to leave this state.

Then by definition the expectation of X is given by:

$$E(X) = \sum_{k=1}^{k=+\infty} k * p^{k-1} * (1-p) = \frac{1}{1-p} \tag{2}$$

Consequently:

$$p = \frac{E(X) - 1}{E(X)} \tag{3}$$

The state transition probabilities can be estimated by means of equation 3.

Next the Viterbi decoding algorithm was applied to each utterance to determine an optimal sequence of states. Consequently, a state of the HMM model was associated to each parameter vector. Each vector was then compared to all the 4 Gaussian mean vectors of the related state —using an Euclidean distance— and was formally associated to the nearest one. When this stage was completed I got 4 classes of vectors. A k-means algorithm was applied to these classes and after this classification process, the new mixtures were updated. The model parameters were re-estimated repeatedly until the estimates remained fixed or the maximum number of iterations was reached. For my case the maximal number of iterations was 20.

3. CALCULATION OF THE DERIVATIVES

Let be $x = x(1), x(2), ...x(k-1), x(k), x(k+1), ...x(I)$ the stream of primary Mel-cepstrum vectors. I calculate the derivative x' of x as follows:

$$x' = x(2) - x(1), x(3) - x(1), x(4) - x(2), ..., \\ x(k+1) - x(k-1), ..., x(I) - x(I-2), x(I) - x(I-1) \tag{4}$$

Equation 4 applied to x' gives the second derivative x'' and so on...

Instead of using equation 4, a most popular linear regression derivative formula can be used [7]. In my application poorer recognition scores than those obtained by means of equation 4 were produced. A potential explanation is that linear regression formulae smooth context information. That seems to be the reason why delta derivatives whose order are greater than 2 are not used is all systems implemented up to now: they are noisy and they do not improve actually system performances. On the contrary, high order derivatives are not to much noisy. They contain useful information for clean speech recognition. The experiments I conducted confirm this point of view.

The primary Mel-cepstrum vectors used were composed of the log energy and the first 11 Mel-cepstrum coefficients. The derivatives of these vectors were calculated by equation 4 and concatenated to the primary vector.

4. EXPERIMENTS

4.1. The Database

I used for my experiments the database Isolet of OGI [1]. This database contains the English alphabet letters produced by 75 male and 75 female speakers. Each speaker uttered each letter twice. Thus, this database contains 7800 utterances. They are divided in 5 groups: Isolet1, Isolet2, Isolet3, Isolet4 and Isolet5. For carrying out my experiments I used 4 groups for training and the fifth group was used for testing. Consequently, 5 tests were possible with this database. Each group has an equal number of speakers. The sampling frequency for this database is 16 khz.

Most speech files of the Isolet database are correctly endpointed, but some files are not well segmented. The performances of a recogniser can therefore be improved if the files are automatically segmented by a good endpoint algorithm. I used the segmentation provided with the database because I did not have access to such an algorithm. Consequently, it was actually a challenge to reach Karnjanadecha's results in these conditions.

4.2. Acoustical Parametrisation

All the speech Mel-cepstrum vectors were calculated on a windowed signal of 32 ms. This temporal window was shifted every 8ms. The c_0 cepstrum coefficient was discarded. The log energy was calculated on a 32 ms portion of the speech signal.

4.3. Experiment I

The influence on recognition accuracy of the successive derivatives was analysed in the first experiment. I firstly incorporated in the primary Mel-cepstrum vector the first two derivatives, then the first three , etc... up to the first six. For this test Isolet1, Isolet2, Isolet3 and Isolet4 were used for training and Isolet5 for testing. Table 1 depicts the corresponding recognition scores.

Derivatives	Recognition accuracy (%)
1st and 2nd Derivative	95.96
+3rd Derivative	96.09
+4th Derivative	97.05
+5th Derivative	97.37
+6th Derivative	97.24

Table 1. Results with the incorporation of successive derivatives in the feature vectors.

The overall performance increased monotonically by adding successive derivatives up to the fifth one. The inclusion of the third, fourth and fifth derivatives increased the recognition score by 1.41% which is not at all negligible. In terms of number of errors this means that with the first two derivatives 63 errors were made instead of 41 for the first five. The results therefore show the benefit of incorporating the successive derivatives in the parameter vectors.

4.4. Experiment 2

The second experiment was conducted in order to test the overall performance of the system. The testing set was rotated, i.e., each set was used as a test set and the remaining ones were used as training sets. I included the first five derivatives in the feature vectors because the first experiment showed no improvements in the recognition scores with other derivatives. Table 2 compares the results published by Karnjanadecha et al. [5] with the High Order Derivatives method (**HOD**):

Testing set	[5](%)	HOD (%)
	Files endpointed	Files not endpointed
Isolet1	97.9	97.05
Isolet2	97.4	97.95
Isolet3	97.4	97.31
Isolet4	97.4	98.01
Isolet5	98.0	97.37
	Average=97.6	Average=97.54

Table 2. Results obtained with different test sets.

My best result was obtained with Isolet4 as testing set: 98.01%. The confusion matrix concerning this test is given in Figure 2. The recognition rate of alphabet letters in the E-set is: 95.77% and for the (m,n) pair I obtained: 96.66%. This latter score is quite remarkable. In all my tests the worst recognised letter of the E-set was letter B which was often confused with letters D, E , P , T or V.

Taking the statistical confidence interval (for this task 0.3%) into account, results produced by means of the two methods can be considered as identical.

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z	Letter	Accuracy
A	60	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	A	100,00
B	0	55	0	1	1	0	0	0	0	0	0	0	0	0	0	3	0	0	0	0	0	0	0	0	0	0	B	91,67
C	0	0	59	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	C	98,33
D	0	2	0	56	1	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	D	93,33
E	0	2	0	0	58	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	E	96,67
F	0	0	0	0	0	59	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	F	98,33
G	0	0	0	0	0	0	59	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	G	98,33
H	0	0	0	0	0	0	0	60	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	H	100,00
I	0	0	0	0	0	0	0	0	60	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	I	100,00
J	0	0	0	0	0	0	0	0	0	60	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	J	100,00
K	0	0	0	0	0	0	0	0	0	0	60	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	K	100,00
L	0	0	0	0	0	1	0	0	0	0	0	59	0	0	0	0	0	0	0	0	0	0	0	0	0	0	L	98,33
M	0	0	0	0	0	0	0	0	0	0	0	0	57	3	0	0	0	0	0	0	0	0	0	0	0	0	M	95,00
N	0	0	0	0	0	0	0	0	0	0	0	0	1	59	0	0	0	0	0	0	0	0	0	0	0	0	N	98,33
O	0	0	0	0	0	0	0	0	0	0	0	0	0	0	60	0	0	0	0	0	0	0	0	0	0	0	O	100,00
P	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	60	0	0	0	0	0	0	0	0	0	0	P	100,00
Q	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	60	0	0	0	0	0	0	0	0	0	Q	100,00
R	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	60	0	0	0	0	0	0	0	0	R	100,00
S	0	0	0	0	0	2	0	0	0	0	0	0	0	0	0	0	0	0	58	0	0	0	0	0	0	0	S	96,67
T	0	0	0	0	0	0	2	0	0	0	0	0	0	0	0	0	0	0	0	58	0	0	0	0	0	0	T	96,67
U	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	60	0	0	0	0	0	U	100,00
V	0	4	0	1	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	54	0	0	0	0	V	90,00
W	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	60	0	0	0	W	100,00
X	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	60	0	0	X	100,00
Y	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	60	0	Y	100,00
Z	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	58	Z	96,67

Fig. 2. The confusion matrix corresponding to my best result (2% of error rate). Isolet 4 was the testing set.

5. COMPARISON BETWEEN THE KARNJANADECHA AND HOD SYSTEMS

First of all, the two systems differ in their HMM recognisers. The Karnjanadecha's system is based on the Cambridge University HTK toolkit [6]. The HOD system is based on my own HMM recogniser. The two systems are comparable in principle, but they are different.

Another important difference between the two systems concerns the feature vector parametrisation techniques used:

- Karnjanadecha et al. [4] [5] manage blocks of cepstral coefficients of variable lengths. For each block they apply a Kaiser windowing and after that a Discrete Cosine Transform (DCT). They retain only 5 coefficients from the DCT. This process is done repeatedly for each of the 10 features of the primary Mel-cepstrum vector. Consequently, after parametrisation, the vectors contain 50 spectral/temporal features. The length of the blocks is variable. The length of the first block is 6 frames which corresponds to 45 ms. The

length increases until it reaches 40 frames, which corresponds to 215 ms. This maximal length is kept to this value until the end of the word is reached. At this point the length of the blocks is gradually reduced until it reaches 6 frames. The Kaiser window beta parameter is also variable. At the beginning of the word its value is 0, then it increases to 5 for blocks whose length is equal to 40 frames. Thus, the features gave better time resolution at the onset and offset portions of each word and less time resolution in other portions of each word.

- I achieved the same results as Karnjanadecha's ones by just incorporating successive derivatives. I simply extended the very popular delta computation paradigm which has been successfully applied to difficult problems like phone recognition [8]. My technique is therefore not just limited to isolated word recognition problems. My technique is a general parametrisation method that does not need to manage blocks of frames of variable length. The HOD technique should therefore be applied successfully to other difficult tasks like connected letter or phone recognition.

On one hand, the methodology proposed by Karnjanadecha et al. [5] for extracting the parameter features is an *ad hoc isolated word recognition based technique*. On the other hand, the advantage of the Karnjanadecha's method compared to mine, is that it requires less computation time and memory, because of the lower dimension of their feature vectors (50 instead of 72 for me).

6. CONCLUSION AND PERSPECTIVE REMARKS

I have proposed in this paper a new methodology for high performance alphabet recognition. My method consists in concatenating successive derivatives to create the feature vectors. Up to now only the first two delta derivatives have been used in speech recognition systems. But the results presented here show that the third, the fourth and the fifth derivative are useful for alphabet letter recognition. These results are as good as the best ones published on alphabet recognition for the same database. My method is quite general and can be applied to other tasks, like for example phone recognition. Kai-Fu Lee and al. [10] were one of the first who tackled phone recognition using what they called **differential coefficients**. Another important point to keep in mind concerning my study is that high order derivatives are not as noisy as delta derivatives and that they contain pertinent information for clean speech recognition.

The good recognition scores produced using the HOD technique show that context information is of capital importance in automatic speech recognition. I think that more powerful context retrieval information algorithms will be necessary for the future for successfully resolving difficult tasks like for example connected alphabet letter or phone recognition.

7. ACKNOWLEDGEMENT

The author would like to acknowledge Dr. Michel Pitermann and Murat Deviren for their help in the writing of this paper.

8. REFERENCES

[1] R. Cole, Y. Muthusamy, M. Fanty, "**The Isolet spoken letter database**" Tech. Rep. 90-004, Oregon Graduate Inst., 1990.

[2] P.C. Loizou and A.S. Spanias, "**High-Performance Alphabet Recognition**", IEEE Transactions on Speech and Audio Processing, Vol. 4, no. 6, pp. 430-445, 1996.

[3] S. A. Zahorian, P. Silsbee, Xihong Wang, "**Phone Classification with Segmental Features and A Binary-Pair Partitioned Neural Network Classifier**", in Proc. Icassp97, April 1997, pp. 1011-1014.

[4] M. Karnjanadecha and S.A. Zahorian, "**Robust Feature Extraction for Alphabet Recognition**", in Proc. ICSLP'98, Sydney, Australia, 1998, pp. 337-340.

[5] M. Karnjanadecha and S.A. Zahorian, "**Signal Modelling for Isolated Word Recognition**", in Proc. Icassp99, pp.293-296.

[6] S.J. Young, J. Odell, D. Ollason, V. Valtchev, P. Woodland, "**Hidden Markov Toolkit V2.1 Reference Manual**", Technical Report, Speech group, Cambridge University Engineering Department, March 1997.

[7] A.C. Surendran, "**Hierarchical Bayes Approach to Adapting Delta and Delta-Delta Cepstra**", in Proc. Icassp 2000, pp. 973-976.

[8] L.F. Lamel and J.L. Gauvain, "**High Performance Speaker-Independent Phone Recognition Using CDHMM**" in Proc. Eurospeech-97, Rhodes, Greece, 1997.

[9] M.J. Hunt, C. Lefebvre, "**A Comparison of Several Acoustic Representations for Speech Recognition with Degraded and Undegraded Speech**", in Proc. Icassp89, Glasgow, Scotland, May 1989, pp. 262-265.

[10] Kai-Fu Lee and Hsiao-Wuen Hon, "**Speaker-Independent Phone Recognition Using Hidden Markov Models**", IEEE Transactions on Acoustics, Speech, and Signal processing, Vol. 37, no. 11, pp. 1641-1648, November 1989.