CALCULATION OF DISTANCE MEASURES BETWEEN HIDDEN MARKOV MODELS

Markus Falkhausen, Herbert Reininger and Dietrich Wolf

Institut für Angewandte Physik, Johann Wolfgang Goethe Universität, FRG
Robert-Mayer-Straße 2-4, D-60054 Frankfurt am Main
markus@apx00.physik.uni-frankfurt.de

ABSTRACT

This paper investigates two methods to define a distance measure between any pair of Hidden Markov Models (HMM). The first one is the geometricaly motivated euclidean distance which solely incorporates the feature probabilities. The second mesures is the Kulback-Liebler distance which is based on the discriminating power of the probability measure on the space of feature sequences induced by the HMMs. A method is shown, to compute the proposed measures reasonable fast and the distance measures are compared in a series of simulations involving HMMs from a real world speech recognition system.

1. INTRODUCTION

Hidden Markov Models have been applied in various research fields. Their success is mainly based upon the existence of a automatic iterativ learning algorithm [1,6] which adjusts the parameter of a HMM to a given training sequence. However, there is no canonical way to measure the dissimilarity between two different HMM. The need for such a distance measure arises in an automatic speech recognition system described in [8], where - according to the current noise situation - distinct sets of HMMs are used for classification. Other applications regard the monitoring of the training procedure, the use of discriminative HMMs, the prediction of expected classification behaviour and the interactive extension of the vocabulary by a user.

Although various meaningful distance measures have been proposed in literature, it has been difficult to calculate them numerically. Here we propose new computational attractive methods for exact and approximative calculations.

The paper is organised as follows, in Paragraph 2 we describe different distance measures and propose various methods to calculate them in Paragraph 3. In the last Paragraph we investigate the speed of convergence for the proposed methods and compare them with results obtained by approximation.

2. DEFINITION OF DISTANCE MEASURES

A HMM $\lambda = (A, B)$ consists of a first order Markov chain with N states from a finite set $S = \{s_1, \dots, s_N\}$ and N related random variables over a common feature space Y. The parameters which define the HMM are the transition probabilities

$$A = (a_{ij}) = (P(S_t = j | S_{t-1} = i))_{i,j = 1, \dots, N}$$

of the Markov chain and the statistics of the random variables

$$B = (b_i(Y)) = (P(Y_t = y | (S_t = i)))_{i = 1, \dots, N} .$$

If the feature space is discrete, i.e. $Y = \{y_1, \dots, y_M\}$, the $(b_i(Y))$ can be written in the form of a stochastic matrix $B = (b_{ik})$.

The probability to observe the state sequence $\underline{S} = (S_1, \dots, S_T)$ and the feature sequence $\underline{Y} = (Y_1, \dots, Y_T)$ is given by

$$P_\lambda(\underline{Y}, \underline{S}) = \prod_{t=1}^{T} a_{S_{t-1}, S_t} b_{S_t}(Y_t) \qquad (1)$$

and the probability of $\underline{Y}$ alone is obtained by summing over all possible state sequences

$$P_\lambda(\underline{Y}) = \sum_{\underline{S}} P_\lambda(\underline{Y}, \underline{S}) . \qquad (2)$$

A distance measure for HMMs is a function which assigns any pair of Models $\lambda = (A, B)$ and $\lambda' = (A', B')$ a positive real number $d(\lambda, \lambda')$. A simple distance measure for discrete HMM is the euclidean distance between the rows of the B-Matrix:

$$d_{ec}(\lambda, \lambda') = \sqrt{\frac{1}{N} \sum_{i=1}^{N} \sum_{k=1}^{M} \| b_{ik} - b'_{ik} \|^2} \quad ,$$

where the number of states of both models has to be identical, i.e. $N = N'$. A more general case is the minimised euclidean distance

$$d_{mec}(\lambda, \lambda') = \sqrt{\frac{1}{N} \sum_{i=1}^{N} \min_j \sum_{k=1}^{M} \| b_{ik} - b'_{jk} \|^2} .$$

However, both measures don‘t take into account the temporal structure represented in the Markov chain and it is easy to find some HMM λ and λ' where the distance attends zero but the generated probability measures P_{λ} and $P_{\lambda'}$ are completely different.

A different distance without this drawback was originally proposed in [4] and measures the discriminative power of P_{λ} and $P_{\lambda'}$ over the feature space Y. The Kullback-Leibler distance [7]

$$d_{KL}(\lambda, \lambda') = \int_{\underline{Y}} \frac{1}{G(\underline{Y})} \log\frac{P_{\lambda'}(\underline{Y})}{P_{\lambda}(\underline{Y})} P_{\lambda}(\underline{Y})\, d\underline{Y} \tag{3}$$

is the expected information gain. $G(Y)$ is a Function which is related to the length $T(Y)$ of the sequence Y. There are two possible definitions which are investigated here. The first one sets G equal to T

$$G_1(\underline{Y}) = T(\underline{Y}) .$$

The second one is a constant which is equal to the expected length of Y

$$G_2(\underline{Y}) = E(T(\underline{Y})) .$$

The first definition will result in a “wordwise” distance measure while the second one will measure the discriminating power “per unit of time”. Note that for defintion in (3) to be meanginful, it is necessary that the probability that a sequence has length greater or equal a given value t tends to zero as t goes to infinity. This is easily achieved if one introduces a final nonemitting stop state [3] in the HMM which signals the end of the sequences.

The KL-distance is not symmetric, i.e.

$$d_{KL}(\lambda, \lambda') \neq d_{KL}(\lambda', \lambda)$$

for some λ, λ'. This feature is appropriate since it is possible that one model can match all typical sequences of the other model rather well, while the other model can‘t do this with all typical sequences of the first model.

3. CALCULATION OF DISTANCES

An analytical solution for the integral in (3) can not derived in general. Even a numerical evaluation has a high computational complexity. We introduce a Monte-Carlo method to evaluate the integral. For this we created a ordered set $F = \{\underline{Y}_1, \ldots, \underline{Y}_L\}$ of feature sequences with a random number generator. Any sequence $\underline{Y}$ could become the l-th member of the set with a probability proportional to $P_{\lambda}(\underline{Y})$. The expected value of the difference of the logarithm of the model scores divided by $G(Y)$ converges to $d_{KL}(\lambda, \lambda')$ as L approaches infinity. Three methods to generate the random numbers were investigated. For the first we used a simple congruential generator described in reference [9]. The second method filtered the random numbers given by the first method to give only typical sequences, where a typical sequence is defined as having relative ocurences of all possible values exactly equal to the expected number of ocurrences. The last method used a gridlike exhausting method of the feature space with a quasi-random sequence [9]. In Figure 1 all three methods are used to fill the unit intervall. As one can see both random methods (left and middle) don’t tend to fill the space equally. In contrast the quasi-random sequence (right) does a more efficient exhaustion of the space.

In addition to (3) we propose as distance measure

$$d_{Vit}(\lambda, \lambda') = \int_{\underline{Y}} \frac{1}{G(\underline{Y})} \log\frac{P_{\lambda'}(\underline{Y}, \underline{S}'_{opt})}{P_{\lambda}(\underline{Y}, \underline{S}_{opt})} P_{\lambda}(\underline{Y})\, d\underline{Y} . \tag{4}$$

which uses the Viterbi-scores

$$P_{\lambda}(\underline{Y}, \underline{S}_{opt}) = \max_{\underline{S}} P_{\lambda}(\underline{Y}, \underline{S}) . \tag{5}$$

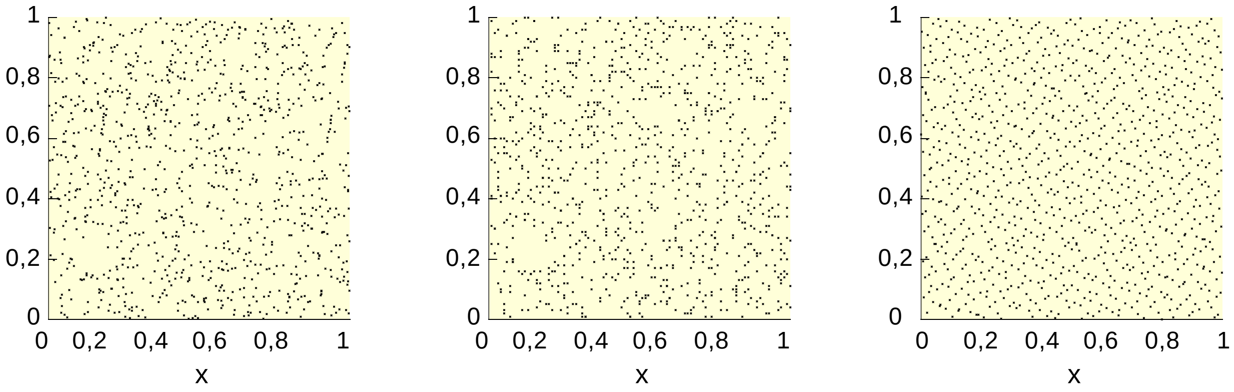

Figure 1: Exhausting the unit square $[0,1]^2$ with Monte-Carlo sequences.
Left: Random number generator.
Middle: Random number generator with typical statistics.
Right: Quasi random sequences.

instead of (2). This approach has the advantage that in many cases the Viterbi-scores are more easily evaluable.

For discrete HMM with the same number of states, a further approximation is possible if $G_2(\underline{Y})$ is used and one assumes that the Viterbi sequences are equal to the state sequence which generated $\underline{Y}_l$

$$\underline{S}'_{opt} = \underline{S}_{opt} = \underline{S}_l .$$

This assumption is reasonable if the two HMM aren't to dissimiliar and the states are discriminating against each other. If the Markov chain is ergodic, equation (4) can be approximated with the ergodic theorem [5]

$$d_{vit}(\lambda, \lambda') = \frac{1}{T(\underline{Y})} \log \frac{P_{\lambda'}(\underline{Y}, \underline{S}_1)}{P_{\lambda}(\underline{Y}, \underline{S}_1)} + \varepsilon$$

with a sufficient long $\underline{Y}$. Note that any HMM with a final stop state can be made ergodic if one connects the final state with the initial states. Therefore a single sequence can be used, instead of a number of sequences. By inserting (1) the last equation results in

$$d_{vit}(\lambda, \lambda') - \varepsilon = \frac{1}{T(\underline{Y})} \sum_{t=1}^{T-1} \log a'_{s_t, s_{t+1}} - \log a_{s_t, s_{t+1}} +$$

$$\frac{1}{T(\underline{Y})} \sum_{t=1}^{T} \log b'_{s_t, y_t} - \log b_{s_t, y_t} .$$

By counting all the instances where $S_t = i, S_{t+1} = j$ or $S_t = i, y_t = k$ for all possible combinations of (i, j) or (i, k) we obtain

$$d_{vit} = \sum_{i,j} \frac{|\{t | S_t = i, S_{t+1} = j\}|}{T(\underline{Y})} (\log a'_{i,j} - \log a_{i,j}) +$$

$$\sum_{i,j} \frac{|\{t | S_t = i, y_t = k\}|}{T(\underline{Y})} (\log b'_{i,k} - \log b_{i,k})$$

If $\underline{Y}$ is long enough it is reasonable to assume that

$$\frac{|\{t | S_t = i, S_{t+1} = j\}|}{T(\underline{Y})} \rightarrow a_{i,j} P_{\lambda}(s_i)$$

and

$$\frac{|\{t | S_t = i, y_t = k\}|}{T(\underline{Y})} \rightarrow b_{i,k} P_{\lambda}(s_i) \qquad .$$

As final result we receive

$$d_{vit}(\lambda, \lambda') \approx \tilde{d}_{vit} = \sum_{i,j} a_{ij} P_{\lambda}(s_i) (\log a'_{ij} - \log a_{ij}) +$$

$$\sum_{i,k} b_{ik} P_{\lambda}(s_i) (\log b'_{ik} - \log b_{ik}) . \qquad (6)$$

The probabilities $\pi_i = P_{\lambda}(s_i)$ can be obtained by solving the linear equation $\pi = \pi A$. Equation (6) is a way to compute the KL-distance as fast as the euclidean distance.

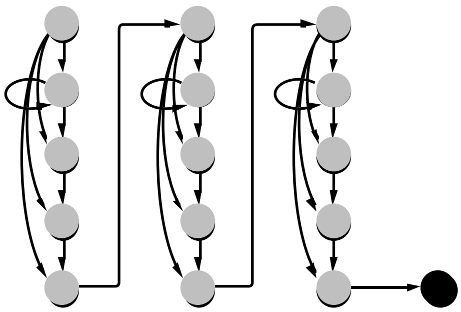

Figure 2: Internal state duration model

4. SIMULATION RESULTS

For numerical evaluation we used 3 isolated word models from our speech recognition system. The HMM had been trained with 50 speakers. All HMMs used an internal state duration model [2] depicted in Figure 2. All states in the same column use the same probability functions for the elements of the feature space, therefore the number of parameters doesn't increase with the number of vertical states. With the transition probabilities between the states in the vertical chain one can exactly model all state durations up to a given maximum. In contrast a standard HMM has only a geometrical distribution for the state durations, where the shortest sequences are the most likely ones. The state duration modelling is a critical feature in the calulation of distance measures, because it assures that the generated sequences for the evaluation of (3) have all a reasonable middle length. Distances between HMMs without internal durational modelling become meaningless, because the generated sequences have very different lengths from the sequences of real speech data.

First we tested the speed of the Monte-Carlo algorithm. Therefore we generated 14 families of sets F_{ν} of feature sequences, each set with $2^{\nu-1}$ members. For every Family we calculted the KL-Distance. Figure 3 shows the performance for every Monte-Carlo generator from Figure 1. The best convergence is achieved with the quasi-random sequence. For a good precision the length of the sequence should become larger then 6000. This is equivalent to a size of the family of 256 members.

Figure 4 displays the logarithmic difference between the results of a set with 8192 members against the logarithm of the total length of the 13 other Families, if a quasi-random generator is used. The solid line in

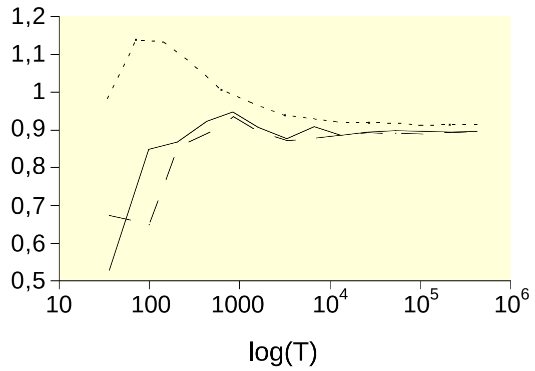

Figure 3: Convergence of random (pointed), typical (dashed) and quasi-random (solid) Monte-Carlo sequences

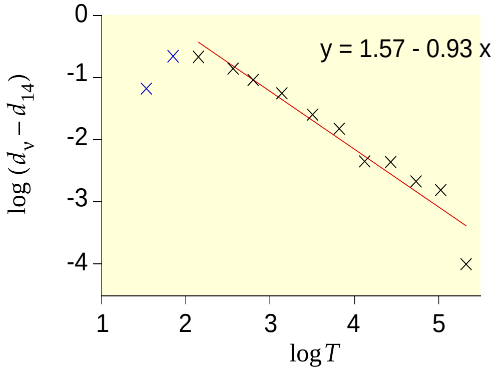

Figure 4: Speed of convergence for a quasi-random sequence

the figure is a regression line with a slope of 0.93. Therefore the precision of our distances goes approximately with $T^{0,9}$. This result assures the computability of the KL-measure for HMMs used in real applications.

In Table 1 the different HMM distance measures are compared to each other. We used a model for the german digit “Neun” as reference and calculated the distance to the phonetically similar word “Nein”, the digits “Null” and “Eins” and the very dissimilar words “Zwo” and “Acht”. The automatic speech recognition system described in [2,8] confused these words in this order. First, one observes that in all cases the distance to the more dissimilar words is higher. No difference is found between the euclidean and the minimised euclidean distance or the use of the word length functions $G_1(Y)$ or $G_2(Y)$. Only minor differences are observed between the KL-distance and the Viterbi-distance. This is consistent with the well known fact that the Viterbi-scores in (5) have nearly identical behaviour to (1). One observes further that equation (6) approximates the Viterbi-distance quite well especially for siniliar words.

Table 1: Distance Measures

	Nein	Null	Eins	Zwo	Acht
d_{ec}	**0.11**	**0.14**	**0.18**	**0.24**	**0.27**
d_{mec}	0.11	0.14	0.18	0.24	0.27
d_{Vit}, G_1	0.63	0.90	1.21	1.50	1.87
d_{Vit}, G_2	0.63	0.90	1.20	1.49	1.87
d_{KL}, G_1	0.64	0.91	1.21	1.51	1.89
d_{KL}, G_2	0.63	0.90	1.21	1.50	1.88
$\tilde{d}_{vit}$	0.71	1.04	1.95	1.94	2.00

5. CONCLUSION

We introduced various distance measures and showed methods to compute them. The distances were compared in a set of simulation experiments. An efficiently computable approximation is shown to perform well with real speech data.

6. REFERENCES

[1] L. Baum, T. Petrie, G. Soules and G. Weis, “A Maximisation Technique Occurring in the Statistical Analysis of Probabilistic Functions of Markov Chains”, An. Math. Stat., 1970, pp. 164-71

[2] M. Falkhausen, N. Nicol, H. Reininger and D. Wolf, “Modellierung von Verweildauern in Hidden-Markov-Modellen”, DAGA 93, Bad Honnef 1993, pp. 977-80

[3] M. Falkhausen, H. Reininger, D. Wolf, “Modellierung der zeitlichen Abfolge von phonetischen Zuständen bei automatischer Spracherkennung mit Hidden-Markov-Modellen.” Informatik Fachberichte 253. Heidelberg 1990, 264-269

[4] B.-H. Juang and L. Rabiner, “A Probabilistic Distance Measure for Hidden Markov Models”, AT&T Technical Journal, Feb. 1985, pp. 391-408.

[5] P. R. Halmos, “Lectures on Ergodic Theorie”, Chelsea, New York 1956

[6] S. Levinson, L. Rabiner and M. Sondhi, “An Introduction to the Application of the Theory of Probabilistic Functions of Markov Process to Automatic Speech Recognition”, Bell Sys. Technical Journal, April 1983, pp. 1035-74

[7] S. Kullback, “Information Theory and Statistics”, Dover Publ., Dover 1968.

[8] N. Nicol, S. Euler, M. Falkhausen, H. Reininger and D. Wolf, ”Noise Classification using Vector Quantisation”, Proc. Eusipco, Edinburgh 1994, pp. 1705-08.

[9] W. Press, S. Teukolsky, W. Vetterling and B. Flannery, “Numerical Recipes in C”, Cambridge Univ. Press, Cambridge 1992.