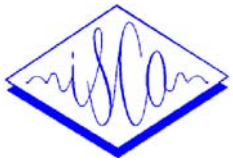

6th European Conference on
Speech Communication and Technology
(EUROSPEECH'99)
Budapest, Hungary, September 5-9, 1999

# A MLE Algorithm for the K-NN HMM System

*Fabrice Lefèvre, Claude Montacié and Marie-José Caraty*

Laboratoire d'Informatique de Paris VI
4, place Jussieu – 75252 PARIS Cedex 05
e-mail : Fabrice.Lefevre@lip6.fr



## Abstract

In this paper, a theoretical framework is proposed for the introduction of the K-NN pdf estimator in an HMM-based speech recognition system. The estimation of the state output probabilities with the K-NN pdf estimator is shown to imply the introduction of a new parameter : the membership coefficient. To learn this coefficient with the Baum-Welch algorithm, a maximum likelihood (ML) reestimation formula is derived. This new formula is tested and compared with the formula we had introduced previously [1].
Then, the edition/condensation techniques are introduced in the context of Markov models in an attempt to improve the appropriateness of the reference data set to the K-NN HMM system. Two new algorithms are proposed for editing and condensing the reference set which present the advantage of being compatible with the K-NN rule.

## Introduction

Hitherto, the theoretical foundations for the introduction of the K-Nearest Neighbours (K-NN) probability density function (pdf) estimator in the Hidden Markov Model (HMM) framework had not been introduced. The first experiments were presented through an empirical transposition of the standard HMM techniques [1]. In this paper, we propose a formal framework for the application of the K-NN estimator in the Markov models. Two points are studied : -the state output probability (outprob) based on the K-NN estimator which requires the introduction of a new parameter (the membership coefficient) and -the reestimation formula for the Baum-Welch training algorithm. Consequently, a procedure for the first estimates of the membership coefficients is proposed.
In a second part, the standard edition and condensation techniques are introduced and tested within the K-NN HMM system. In our idea, these techniques could help to increase the efficiency of the reference set used with the K-NN estimator in the HMM system. In that purpose, two new algorithms are proposed : -a new way of editing a data set based on the behaviour of the points and -a new condensation criterion taking account of the level of identification. These new methods have the advantage of being entirely based on the K-NN whereas usual algorithms rely on the 1-NN.

## 1 MLE in the K-NN HMM System

The first step of the formal introduction of the K-NN estimator in the HMM framework is to re-formulate the outprobs based on the K-NN pdf estimator. Let's recall that a first order S-states HMM is defined by a SxS transition matrix $\{a_{ij}\}$ and S pdfs $\{b_i\}$ generating the outprobs. For sake of simplicity, the initial probability vector is not mention and will be considered 1 on the initial state and 0 on the others.

### *1.1 Membership Coefficients*

With the K-NN estimator, the outprob of state $i$ for the vector $x_0$ can be basically expressed, after having computing its K-NN from the reference data set, as :

$$b_i(x_0) = k_i / n_i \qquad (1)$$

where $k_i$ is the number of $x_0$'s nearest neighbours associated with state $i$ and $n_i$ is the total number of reference points associated with state $i$. This raises the issue of associating the reference points to the HMM states.
In case of a phonetically labelled database, the reference points can be straightforwardly associated to the HMM of their phonetic class. But, thereafter, the association at the state level should be automatically obtained since no expertise is available at that level.
A smooth assignment of the reference points to the states is preferable to an one-to-one association as it will maintain more information. Smoothness could be represented by a degree of membership. Every reference point $x_0$ will be assigned membership coefficients $u_i$ on every state $i$ :

$$u_i(x_0) \in [0,1] \quad and \quad \sum_{i=1}^{S_G} u_i(x_0) = 1 \qquad (2)$$

with $S_G$ the global number of distinct states in the system. Then, the outprobs for $x_0$ on state $i$ can be expressed as :

$$b_i(x_0) = \frac{\sum_{v=1}^{V} u_i(v)\, nn_K(x_0, v)}{U_i} \quad \text{with } U_i = \sum_{v=1}^{V} u_i(v) \qquad (3)$$

with $V$ the number of reference points and $nn_K(x_0,v)$ is a function which is 1 if $v$ belongs to the K-NN of $x_0$, 0 otherwise :

$$nn_K(x_0, v) = \begin{cases} 1 \text{ if } v \in \text{K-NN of } x_0 \\ 0 \text{ otherwise} \end{cases} \qquad (4)$$

The membership coefficients can serve directly to improve the classification based on the K-NN rule [2, 3].

In this case, their values are generally obtained from distance considerations. In our case, they will be learned by means of the MLE-based Baum-Welch algorithm.

### *1.2 Reestimation Formula for MLE Training*

The parameters of the K-NN estimator are the membership coefficients of the reference points. A new ML reestimation formula suitable for the Baum-Welch algorithm is derived. The following demonstration borrows heavily to [4] where the reestimation formulae were derived for the mixture of gaussians pdf estimator. Our main contribution lies in the introduction of one more dimension in the Cartesian product defining the paths along an HMM : the nearest neighbours sequence.

The likelihood of an acoustic segment $O = o_1,..,o_T$ over an HMM is obtained via the integration over the hidden state sequences :

$$Q = q_1,\ldots,q_t,\ldots,q_T \tag{5}$$

The use of the K-NN estimator implies to consider also the nearest neighbours hidden sequences :

$$N = n_1,\ldots,n_t,\ldots,n_T \tag{6}$$

Then, the likelihood of segment $O$ upon model M for the particular sequences $Q$ and $N$ is :

$$P(O,Q,N/M) = \prod_{t=1}^{T} a_{q_{t-1}q_t}\, b_{q_t}^{n_t}(o_t) \tag{7}$$

with the outprob given by :

$$b_{q_t}^{n_t}(o_t) = \frac{u_{q_t}(n_t)\, nn_K(o_t,n_t)}{U_{q_t}} \tag{8}$$

An interpretation of (7) is that it exists $N^T$ state sequences that may lead to the observation $O$ and for each of them $V^T$ possible branch sequence among the reference points due to the K-NN estimate. Practically, only $K^T$ of them are actually encountered. The complete likelihood is obtained by summing (7) over all possible sequences of states and nearest neighbours :

$$P(O/M) = \sum_{Q,N}\left\{\prod_{t=1}^{T} a_{q_{t-1}q_t}\, b_{q_t}^{n_t}(o_t)\right\} \tag{9}$$

The maximisation of (9) is obtained by maximisation of an auxiliary function based on the Kullback-Leibler statistic [5] :

$$F(M,\hat{M}) = \sum_{Q,N}\left\{P(O,Q,N/M)\log P(O,Q,N/\hat{M})\right\} \tag{10}$$

In [4], it is shown that increasing $F$ as a function of $\hat{M}$ will monotonically increase the likelihood of the given observation among the model. The auxiliary function maximisation can be parted in two sub-maximisation due to the log :

$$\log P(O,Q,N/\hat{M}) = \sum_{t=1}^{T}\log \hat{a}_{q_{t-1}q_t} + \sum_{t=1}^{T} nn_K(o_t,n_t)\log\frac{\hat{u}_{q_t}(n_t)}{U_{q_t}} \tag{11}$$

and then :

$$F(M,\hat{M}) = \sum_{i=1}^{S} F_{\hat{a}_i}(M,\{\hat{a}_{ij}\}_1^S) + \sum_{i=1}^{S} F_{\hat{u}_i}(M,\{\hat{u}_i(v)\}_1^V) \tag{12}$$

The first term of (12) will lead to the conventional reestimation formula for the transition matrix elements. The second one is developed :

$$F_{\hat{u}}(M,\{\hat{u}_i(v)\}_1^V) = \sum_{Q,N} P(O,Q,N/M)\sum_{t=1}^{T} nn_K(o_t,n_t)\log\frac{\hat{u}_{q_t}(n_t)}{U_{q_t}}\delta_i^{q_t} = \sum_{v=1}^{V}\sum_{t=1}^{T} P(O,n_t=v,q_t=i/M)\, nn_K(o_t,v)\log\frac{\hat{u}_i(v)}{U_i} \tag{13}$$

with $\delta$ the Kronecker symbol. The general solution for the maximisation of equations alike (13) can be found in [4] and leads to :

$$\frac{\hat{u}_i(v)}{U_i} = \frac{\sum_{t=1}^{T} P(O,n_t=v,q_t=i/M)\, nn_K(o_t,v)}{\sum_{t=1}^{T} P(O,q_t=i/M)\, nn_K(o_t,v)} \tag{14}$$

Finally, a solution to the equation (14) is given by :

$$\hat{u}_i(v) = \frac{\sum_{t=1}^{T} P(O,n_t=v,q_t=i/M)\, nn_K(o_t,v)}{\sum_{t=1}^{T} P(O,n_t=v/M)\, nn_K(o_t,v)} \tag{15}$$

which could be expressed with the standard forward-backward variables $\alpha$ and $\beta$ [6] :

$$\hat{u}_i(v) = \frac{\sum_{t=1}^{T}\sum_{j=1}^{S}(\alpha_{t-1}(j)\,a_{ji})\,u_i(v)\,nn_K(o_t,v)\beta_t(i)}{\sum_{t=1}^{T}\sum_{i,j=1}^{S}(\alpha_{t-1}(j)\,a_{ji})\,u_i(v)\,nn_K(o_t,v)\beta_t(i)} \tag{16}$$

This result can be generalised to a set of $E$ examples :

$$\hat{u}_i(v) = \frac{\sum_{e=1}^{E}\sum_{t=1}^{T_e}\sum_{j=1}^{S}(\alpha_{t-1}^e(j)\,a_{ji})\,u_i(v)\,nn_K(o_t^e,v)\beta_t^e(i)}{\sum_{e=1}^{E}\sum_{t=1}^{T_e}\sum_{i,j=1}^{S}(\alpha_{t-1}^e(j)\,a_{ji})\,u_i(v)\,nn_K(o_t^e,v)\beta_t^e(i)} \tag{17}$$

In our previous papers [1,7], an empirical reestimation formula was used. The principle was to evaluate the membership coefficients of the reference points directly by the probability for this point to be emitted by this state considering every possible alignment of its segment on the model :

$$\hat{u}_i(o_t) = P(O,q_t=i/M)/P(O/M) \tag{18}$$

which could be expressed with the forward-backward variables as :

$$\hat{u}_i(o_t) = \alpha_i(t)\times\beta_i(t)/\alpha_{i_F}(T) \tag{19}$$

where $i_F$ is the final state. This formula can also be generalised to a set of E examples :

$$\hat{u}_i(o_t) = \sum_{e=1}^{E}\alpha_i^e(t)\times\beta_i^e(t)\Big/\sum_{e=1}^{E}\alpha_{i_F}^e(T_e) \tag{20}$$

This formula proceeds locally contrary to (17) which propagates the reestimation on the nearest neighbours. The formula (20) is simpler and faster but has two drawbacks : the reference and training sets should be the same and no convergence proof is established.

## 2 Training the 50-NN HMM system with MLE

The experiments are performed on the phonetically-labelled TIMIT database. Computed each centi-second, a vector-frame is represented by 12 Mel-Frequency Cepstrum Coefficients and by the energy coefficient. The system comprises 39 3-states Bakis HMMs, modelling the phonetic classes. Two core sets are extracted from the main data sets : -the core-test (192 sentences, 7,215 phones and 57,919 frames), -the core-train (192 sentences, 7,381 phones and 57,669 frames). The reference data set used for the 50-NN is composed of the 1,124,823 frames of the entire training set.
The first estimates of the membership coefficients are obtained from a simple procedure : each reference point is given an equal membership coefficient (*1/N*) on every state of its HMM.

### *2.1 Training Convergence*

No convergence proof has been established for the local reestimation. Actually, the experiments show that the procedure reaches local minima but escapes from them right after. To avoid this problem, the epochs are stopped when a negative variation of the average log probability is encountered.

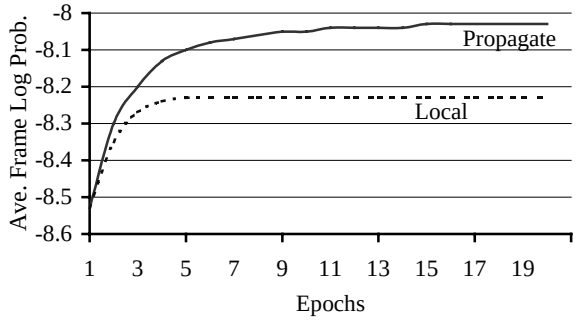


Figure 1 Convergence of the training procedures

In Figure 1, the evolution of the global average frame log probability is represented as a function of the epochs. The global log probability is obtained as the mean of the average frame log probability for every models. The propagate reestimation presents a better behaviour : it reaches a higher global log probability value and continues to progress as the epochs go by.

### *2.2 Acoustic-phonetic Decoding*

The results of the acoustic-phonetic decoding are reported on Table 1 for the local and propagate reestimation trainings. A phonetic bigram, learned on the training set, is used.

| | Reestimation | Ident | Subst | Del | Ins | Acc |
|---|---|---|---|---|---|---|
| Train | Local | 66.6 | 18.0 | 15.5 | 3.6 | 63.0 |
| | Propagate | 69.5 | 16.2 | 14.3 | 3.4 | **66.1** |
| Test | Locale | 58.5 | 25.4 | 16.1 | 4.5 | 54.0 |
| | Propagate | 58.9 | 25.1 | 16.0 | 4.6 | 54.3 |

Table 1 Acoustic-phonetic decoding recognition rates (%) on the core sets of TIMIT with different reestimation formulae

On the test set, the rates can not be statistically distinguished (95% confidence interval radius is 1.1%). The propagate reestimation gives models well-fitted to the training set (which is, in our case, a subset of the reference set). A lack of generalisation ability could explain that the discrepancy is not conserved on the test set. Our assumption is that a better generalisation could be obtained from the K-NN HMM system by defining a reference set more appropriate to the K-NN estimator. In this purpose, the standard edition and condensation techniques are introduced.

## 3 Edited/condensed Data Sets

The edition and the condensation are usual techniques for improving the NN rule. Their properties could be profitable to the K-NN HMM system.
Editing a data set consists in removing the "bad" points whereas the condensation eliminates the useless points. Both procedures result in decreasing the computation cost. But only, the edition is expected to improve the recognition.

### *3.1 Editing the Reference Set*

The principle of the edition procedure [8] is to remove from the reference set all the points leading to identification mistakes. Thus, both discrimination between classes and computation cost can be improved.
The classical Wilson approach, that Devijver [9] proposed as MultiEdit algorithm, removes the points incorrectly identified :

Treatment :
The reference points are identified by the K-NN rule
The points incorrectly identified are removed
Stopping Criterion :
Repeat n times or stop when a given number of points has been removed

The option, we propose, removes the points when they are involved in too many misclassifications :

Treatment :
The reference points are identified by the K-NN rule
If correct identification Then
For all the nearest neighbours of the correct class :
Counter + bonus
Else
Counter - penalty
The points with a negative counter are removed
Stopping Criterion :
Repeat n times or stop when a given number of points has been removed

The bonus is set to 2 and the penalty to 1. Thus, a point is removed when it has been involved in twice as more misclassifications as correct ones.

### *3.2 Condensing the Reference Set*

The principle of the condensation technique [10] is to remove from the reference set the useless points (i.e. the performance stays the same whether these points are or not in the reference set). Removing these points avoids useless distance computations.
Usually, the condensation of a data set is obtained by the Hart's algorithm which is only defined for the 1-NN rule. We propose an algorithm adapted to the K-NN rule.

Our proposal is based on a different characterisation of the condensed points.
The usefulness of a point being hard to determine explicitly we hypothesise that it is linked to the level of identification : better a point is identified, less it should be in a confusion area and, potentially, less it should be useful for the classification. Practically, the points will be progressively removed starting from those having the higher number of nearest neighbours of their class. The algorithm can be sketched :

Treatment :
  For k=K to 1
    All points having k nearest neighbours of their class are removed (removed points can no more be counted as nn for the following points)
  Compute new error rate of frame identification
Stopping Criterion :
  Error rate has increased by a given factor

## 4 Evaluation of the 50-NN HMM System

For the evaluation, 4 reference data sets have been considered : the initial set, the set edited by the Wilson method (Edition1), the set edited by our method (Edition2) and the set condensed by our method. The K-NN HMM system used is conformed to the previous description and is learned with the propagate reestimation. The training and test sets are not modified. The results are given on a frame identification and an acoustic-phonetic decoding.

### *4.1 Frame Identification*

The frame identification rates, as well as the resulting number of frames, are gathered in Table 2 for the 4 reference sets.

| Reference Set | Nb of frames | Train | Test |
|---|---|---|---|
| Initial | 1,1124,823 | 56.8 | 52.6 |
| Edition1 | 645,396 | 54.1 | 52.6 |
| Edition2 | 580,216 | 54.0 | 52.4 |
| Condensation | 905,305 | 55.9 | 51.7 |

Table 2 Frame identification rates (%) on the train and test core sets of TIMIT considering four reference sets

The results for the edition show that the reference set can be halved without noticeable decrease in identification rates. The edition has reduced some phonetic classes (the occlusives mainly) to 10% of their initial sizes. It results in unrecognised classes. But, this effect is compensated by an increase of the scores of the remaining classes. Also, the silence class plays a special role since it is massively conserved at the expense of the others classes (23% in the initial set and 40% in the edited sets). This is an undesirable effect as the condensed set results show that their proportion can be reduced to 10% without a great drop in recognition performance.

### *4.2 Acoustic-phonetic Decoding*

The recognition results on the training and test cores of TIMIT are given in Table 3 for the 4 reference sets. Unlike for the frame identification, the edited sets rates are neatly inferior to those of the initial set. In the case of the acoustic-phonetic decoding, not only the deleted classes can not be recognised but also it implies an increase of the insertion rate.

| | Reference Set | Ident | Subst | Del | Ins | Acc |
|---|---|---|---|---|---|---|
| Train | Initial | 69.5 | 16.2 | 14.3 | 3.4 | 66.1 |
| | Edition1 | 62.5 | 25.6 | 11.9 | 11.7 | 50.7 |
| | Edition2 | 61.7 | 26.1 | 12.2 | 12.2 | 49.5 |
| | Condensation | 56.9 | 18.4 | 24.7 | 3.3 | 53.6 |
| Test | Initial | 58.9 | 25.1 | 16.0 | 4.6 | 54.3 |
| | Edition1 | 57.7 | 30.1 | 12.2 | 10.2 | 47.5 |
| | Edition2 | 57.2 | 30.5 | 12.3 | 10.9 | 46.3 |
| | Condensation | 56.7 | 26.1 | 17.2 | 4.1 | 52.6 |

Table 3 Acoustic-phonetic decoding results (%) on the train and test core sets of TIMIT considering four reference sets

For its part, the condensed set, while clearly declining on the train set falls only of a few per cents on the test set.

## Conclusion and Perspectives

A formal framework for the introduction of the K-NN pdf estimator in the HMM has been presented. A training scheme has been obtained by the introduction of the membership coefficients in the K-NN HMM system and the derivation of a suitable ML reestimation formula.
Some insights have been gained in the use of the edition and condensation techniques in the K-NN HMM system. If the condensation has performed as expected, the edition technique has not lead to the expected results. Our future work will consist in the introduction in the edition procedure of a mechanism for maintaining the proportionality of the phonetic classes so as to reduce the insertion rate of the system.

## References


1. Lefèvre, F., C. Montacié, and M.J. Caraty, *K-NN Estimator in a HMM-Based Recogntion System*, in *Computational Models of Speech*, 1997, NATO ASI.
2. Denoeux, T., *A K-nearest Neighbor Classification Rule Based on Dempster-Shafer Theory.* IEEE Transactions on Systems, Man and Cybernetics, 1995. **25**(5).
3. Keller, J.M., M.R. Gray, and J.A. Givens, *A Fuzzy K-NN Neighbor Algorithm.* IEEE Transactions on Systems, Man and Cybernetics, 1985. **15**(4): p. 580-585.
4. Liporace, L., *Maximum Likelihood Estimation for Multivariate Observations of Markov Sources.* IEEE Transactions on Information Theory, 1982. **28**(5): p. 729-734.
5. Kullback, S. and R. Liebler, *On Information and Sufficiency.* Ann. Math. Statist., 1951. **22**: p. 79-86.
6. Baum, L., *An Inequality and Association Maximization Technique in Statistical Estimation for probabilistic Functions of Markov Processes.* Inequalities, 1972. **3**: p. 1-8.
7. Montacié, C., M.-J. Caraty, and F. Lefèvre. *KNN versus Gaussian in a HMM-Based System*. in *Eurospeech*. 1997.
8. Wilson, D., *Asymptotic Properties of Nearest Neighbor Rules Using Edited Data.* IEEE Transactions on Systems, Man and Cybernetics, 1972. **2**(3): p. 408-421.
9. Devijver, P.A. and J. Kittler, *Pattern Recognition : A Statistical Approach*. 1982. London: Prentice-Hall.
10. Hart, P., *The Condensed Nearest Neighbor rule.* IEEE Transactions on Information Theory, 1968. **14**: p. 515-516.