K-Nearest Neighbours Estimator in a HMM-Based System

Fabrice Lefevre, Claude Montacié, Marie-Josée Caraty

K-Nearest Neighbours Estimator in a HMM-Based Recognition System

Fabrice Lefèvre, Claude Montacié and Marie-José Caraty

LIP6 - Université Pierre et Marie Curie - CNRS
4, place Jussieu - 75252 Paris Cedex 5 - France
e-mail : Lefevre.Fabrice@lip6.fr

Summary. For many years, the K-Nearest Neighbours method (K-NN) has been known as one of the best probability density function (pdf) estimator [2]. The development of fast K-NN algorithms allows to reconsider its use in applications with large sample sets. In this outlook, the K-NN decision principle has been assessed on a frame by frame phonetic identification on the TIMIT database. Thereafter, a method to integrate the K-NN pdf estimator in a HMM-based system is proposed and tested on an acoustic-phonetic decoding task.

Key words: nonparametric estimator; probability density function; HMM training; acoustic-phonetic decoding; SNALC.

1. Introduction

In continuous HMM, the state output distributions are usually represented by Gaussian mixture densities. But most of the time, the analysis vectors do not have a Gaussian distribution [5]. Theoretically, Gaussian mixtures could estimate any pdf. Practically, the number of Gaussian functions in the mixtures is derived from heuristics and its optimality is not warranted. This problem occur with any parametric pdf representing the state output distribution of a HMM. To address this difficulty, we propose to develop a HMM-based system using the nonparametric K-NN pdf estimator.

Basically, the K-NN decision rule assigns to an unclassified sample point the class of the majority of its K nearest neighbours of a set of previously classified points. This method is nonparametric since no assumption is made upon the joint distribution of the sample points. A simple modification of the K-NN rule gives a consistent pdf estimator [4] whose asymptotic performances are well known. The 1-NN error bound is at the most twice the optimal Bayes error, and the K-NN error decreases as K increases. The difficulty raised by the high computational demand of the K-NN estimator has been drastically reduced by the development of fast K-NN algorithms which could avoid up to 99.8% of the systematic computation [6].

2. K-NN Assessment

Two preliminary experiments aim at the assessment of the K-NN estimator. These experiments are carried out on TIMIT database ($1,124,823$ training-frames and $57,919$ test-frames). Computed per centi-second, a frame is represented by 12

MFCC and by the energy coefficient. The TIMIT reference labeling [7] provides the classification of the training-frames.

In the first experiment, for each test-frame, we compute the probability of its expected phonetic class according to a 50-NN estimator and an 8 Gaussian mixture estimator (8-Gaussian). Figure 1 gives the histograms of the expected class probability for all the test-frames.

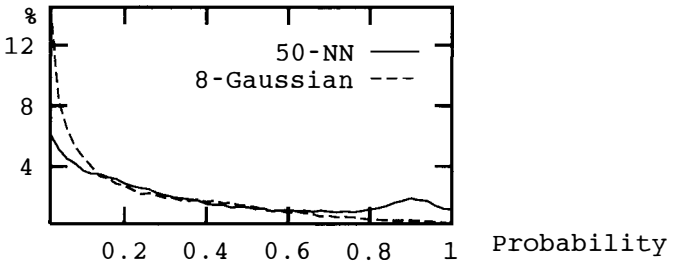

Fig. 1. Histograms of expected class probability for 50-NN and 8-Gaussian estimators

The average expected class probability is 0.36 for the 50-NN and 0.25 for the 8-Gaussian. For the 50-NN, it means that a test-frame has an average of 18 nearest neighbours of the expected class. Thus, three distinctive parts are distinguished. The first one ($k < 18$) represents the vectors far from their proper training set. These vectors have to be studied to analyze precisely the reasons for such spatial distortions. The second part ($k > 25$, i.e. $prob. > 0.5$) represents the vectors which identification is simple with, for instance, the majority rule. The last part ($18 \leq k \leq 25$) represents the vector requiring a more sophisticated identification rule.

In the second experiment, the phonetic identification frame rate has been computed using the 50-NN maximum maximorum decision principle. The global identification rate is about 50%. The explanation of this result is threefold : 1. the low occurrence of phonemes such as [ch, dh, jh, ng, uh, y], 2. the segmentation error inherent to any labeling, 3. the difficulty in identifying complex phonemes such as diphthongs or plosives using a single frame. These statements account for the introduction of the K-NN information in a global decision principle such as HMM.

3. K-NN estimator in HMM

Training and decoding techniques for HMM do not rely on the used pdf. In practice, some adaptations are required from a Gaussian pdf-based system to a K-NN pdf-based system. We have adapted the Forward Backward (FB) and Viterbi algorithms in *HMM ToolKit 1.4* [8] to handle the K-NN estimator.

3.1 Adaptation Principle

The state output probability calculation in HMM is adapted to the K-NN estimator. For this, we compute for each training vector its state occupation probabilities.

The state occupation probability of state s for the vector v of the vector sequence V is computed as

$$P^S_{Occ}(v) = \frac{P(V, X(v) = s|M)}{P(V|M)} \tag{1}$$

where X is a state sequence of V and M the model of V according to a previous or reference alignment. Both numerator and denominator probabilities are derived from the FB or the Viterbi algorithm. The Viterbi algorithm finds the maximum likelihood state sequence, so each observation vector is assigned to a single state. That is, P_{Occ} is 1 for one state and 0 for the others. Whereas, in the FB, the full likelihood is computed on all possible state sequence and each observation vector is assigned to every state in proportion of the model being in that state when the vector was observed.

Thereafter, the state output probability of any vector is calculated as the normalized summation of the state occupation probabilities of its K-NN :

$$P^S_{Out}(v) = \frac{\sum_{k=1}^{K} P^S_{Occ}(k^{th} - NN(v))}{K} \tag{2}$$

where $k^{th} - NN(v)$ is the k^{th} nearest neighbour of observation vector v.

We have introduced this state output probability calculation in the FB and Viterbi algorithms. Usually, the Viterbi algorithm is used to perform a first estimation before the re-estimation with FB. In our case, this first estimation with Viterbi does not fit since nearly all the vectors are finally assigned to one single state. The direct assignment of vectors to individual states in Viterbi creates an irreversible accumulation phenomenon. An original method is proposed to estimate the model parameters without the initial Viterbi stage.

3.2 HMM Estimation Improvement

A HMM training assumes initial estimates of the HMM parameters. Commonly, a rough guess of the initial pdf values is obtained by a uniform segmentation of the observations sequences, associating each successive segment with successive states. Then a first estimation is made with the Viterbi algorithm, which we saw is not adapted in our case. To address this, the first estimation of the models is obtained from a new method using the information provided by the K-NN.

The parameters are still initialized by a uniform segmentation. Then, a time to states projection is performed. That is, each vector is associated to every states proportionally to the number of its K-NN in this state considering the uniform segmentation. Then, the vector sequences are divided into time-proportional intervals.

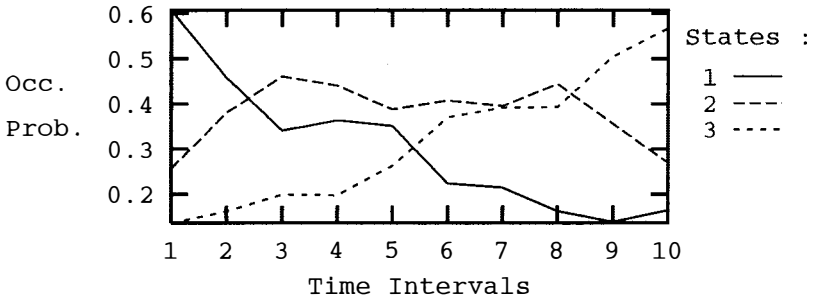

Fig. 2. Time/states projection for vowel /uw/

In this way, the time/states projections are averaged amongst all vectors of each interval. These average time/states projections are used to initialize state occupation probabilities.

Figure 2 illustrates the time/states projection in a three-states Bakis model for the vowel /uw/, with 50-NN. For instance, the state occupation probabilities of the observations of the fifth interval (i.e. included within 40 and 50% of their sequence duration) are 0.26 for the first state, 0.46 for the second one and 0.27 for the third one.

The trainings performed with the time/states projection do not show any significant gain in the accuracy of the model estimation (i.e. do not increase the average model likelihood). Nonetheless, it should be noted that a certain amount of observation sequences could have a null likelihood because of the incoherence between the paths imposed by two successive frames. This effect is rather eliminated using the time/states projection initialization. Thus, the FB convergence should be considered better as it involves more training sequences. The impact on the accuracy rate (+0.2%) is not significant. Others attempts, such as topology inference [1], will be investigated to address the problem of the first estimates of the parameters

4. Evaluations

The experiments aim at comparing the Gaussian and the K-NN estimators in HMM-based system. The chosen task is the reference acoustic-phonetic decoding [3] on TIMIT database core-test. The basis HMM is a three-states Bakis. Both systems use a phonetic back-off bigram learned on the trainset. Two kinds of criterion are used : the recognition rates and the Segmental Normalized Acoustic Likelihood Coefficient (SNALC).

4.1 Recognition rates

Table 1 presents the recognition results for the 8-Gaussian and 50-NN systems. These results are low but are related to the simplicity of the involved models. The gain in identification with the 50-NN estimator is nearly 6%. This difference is

Table 1. Recognition results on TIMIT core-test

			Error Details		
	Identification	Accuracy	Deletion	Substitution	Insertion
50-NN	58.17	51.84	14.91	26.91	6.33
8-Gaussian	52.35	50.06	24.09	23.56	2.29

lowered to 1.8% in accuracy due to the high level of insertion in the 50-NN system (6.3%). This difference between accuracy rates is barely statistically meaningful since the confidence interval for these conditions is 1.1%. To compare the systems thoroughly, we introduce a new criterion : the SNALC.

4.2 SNALC Evaluation

The SNALC is an attempt to obtain a better evaluation of the pdf influence in a decision process such as HMM. This coefficient is computed for each frame during a Viterbi decoding as :

$$SNALC = 1 - \frac{L_{ref}/T_{ref}}{\sum_i L_i/T_i} \quad (3)$$

where L_i is the i^{th} model likelihood for a segment of T_i frames. The reference phoneme corresponds to the TIMIT alignment.

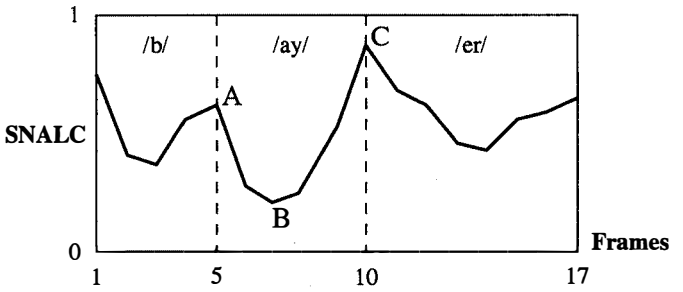

Fig. 3. Typical behaviour of SNALC

Figure 3 is an illustration of the typical behaviour of SNALC, here on an utterance of the world "buyer". From the beginning of the phoneme /ay/ (A), the SNALC decreases to B : the /ay/ model likelihood becomes preponderant. Then, the likelihood of /er/ increases and the SNALC raises up to the end of /ay/ (C). The SNALC values are relevant to confusions between phonemes or co-articulation phenomena. Thus, the average SNALC can be used as an assessment measure.

Experiments were performed on the TIMIT core test in the above-mentioned conditions. The average SNALC for the 50-NN (0.62) is better than the 8-Gaussian one (0.74). If the average SNALC provides a synthetic value for assessment, the

SNALC affords more detailed analysis of the decision process behaviour. Thus, it highlights a resonance effect in the computation of the state output probabilities with the Gaussian estimator : values overstepping 1. Actually, the Gaussian functions are pdf's. They should be integrated around the considered vector to become probabilities. We noted this on SNALC curves for phonetic classes such as /silence/, /f/ or even /s/. This point reveals an uncontrolled effect (the resonance) in the Gaussian HMM-based decision process.

5. Perspectives

For the first time, the K-NN estimator has been introduced in a HMM-based system. At this moment, its performances are comparable to the Gaussian ones. However, it is remarkable that most of the continuous HMM techniques are strongly adapted to Gaussian pdf's. Suitable techniques are to be found for the K-NN estimator.

Improvements of the K-NN HMM-based system could arise from the ability to retrace and to analyze the training vectors causing the identification errors. Another interesting propriety of the K-NN estimator is to provide a general framework in which we could combine information from different representation spaces, and thus use locally adapted spaces and their associated metrics.

Finally, as they behave differently, K-NN and Gaussian should favourably be used simultaneously in a composite pdf estimator.

References

[1] R. de Mori, M. Galler, and F. Brugnara. Search and learning strategies for improving hmm. *Computer Speech and Language*, 9:107–121, 1995.

[2] J. Goût. L'apprentissage en reconnaissance de la parole. Technical report, Université PARIS 6, 1993.

[3] K.-F. Lee and H.-W. Hon. Context-dependent phonetic hmm for speaker-independent continuous speech recognition. *IEEE Trans. ASSP*, 38(4):599–609, 1990.

[4] D. Lotfsgaarden and C. Quesenberry. A nonparametric estimate of a multivariate density function. *Annals Math. Stat.*, 36:1049–1051, 1965.

[5] C. Montacié, M.-J. Caraty, and C. Barras. Mixture splitting technic and temporal control in a hmm-based recognition system. In *Proc. ICSLP*, 1996.

[6] C. Montacié, M.-J. Caraty, and F. Lefèvre. K-NN versus gaussian in a HMM-based recognition system. In *Proc. Eurospeech*, 1997.

[7] S. Seneff and V. Zu. *Transcription and Alignment of the TIMIT Database.* NIST, 1988. CD-ROM TIMIT.

[8] S. Young. *HTK Version 1.4 : Reference Manual and User Manual.* CUED-Speech Group, 1992.