ACADEMIC PRESS

Available online at www.sciencedirect.com

SCIENCE DIRECT®

Computer Speech and Language 17 (2003) 113–136

COMPUTER SPEECH AND LANGUAGE

www.elsevier.com/locate/csl

Non-parametric probability estimation for HMM-based automatic speech recognition

Fabrice Lefèvre [1]

Spoken Language Processing Group, LIMSI-CNRS, BP 133, 91403 Orsay Cedex, France

Received 6 June 2001; received in revised form 5 February 2003; accepted 5 February 2003

Abstract

During the last decade, the most significant advances in the field of continuous speech recognition (CSR) have arisen from the use of hidden Markov models (HMM) for acoustic modeling. These models address one of the major issues for CSR: simultaneous modeling of temporal and frequency distortions in the speech signal. In the HMM, the temporal dimension is managed through an oriented states graph, each state accounting for the local frequency distortions through a probability density function. In this study, improvement of the HMM performance is expected from the introduction of a very effective non-parametric probability density function estimate: the *k*-nearest neighbors (*k*-nn) estimate.

First, experiments on a short-term speech spectrum identification task are performed to compare the *k*-nn estimate and the widespread estimate based on mixtures of Gaussian functions. Then adaptations implied by the integration of the *k*-nn estimate in an HMM-based recognition system are developed. An optimal training protocol is obtained based on the introduction of the membership coefficients in the HMM parameters. The membership coefficients measure the degree of association between a reference acoustic vector and a HMM state. The training procedure uses the expectation-maximization (EM) algorithm applied to the membership coefficient estimation. Its convergence is shown according to the maximum likelihood criterion. This study leads to the development of a baseline *k*-nn/HMM recognition system which is evaluated on the TIMIT speech database. Further improvements of the *k*-nn/HMM system are finally sought through the introduction of a temporal information into the representation space (*delta coefficients*) and the adaptation of the references (mainly, *gender modeling and contextual modeling*).

E-mail address: fabrice.lefevre@limsi.fr.

[1] This work has been carried out in the LIP6 Speech & Audio Indexation Group of the Université Paris VI.

0885-2308/03/$ - see front matter
doi:10.1016/S0885-2308(03)00004-4

1. Introduction

Today, the most efficient CSR systems are based on statistical methods implemented by means of HMMs. Comparing two acoustic realizations of a symbolic unit (such as a phone or a word), the acoustic vector sequences may present complex temporal transformations combined with vector distortions in their representation space (generally the frequency domain). Modeling of acoustic vector sequences attempts to accommodate these two embedded distortion levels. In the statistical recognition framework, the frequency level is dealt with by probability densities which determine the most probable representation space region associated with the a priori classes. For their part, the temporal distortions are handled through dynamic programming. The hidden Markov models provide an ad hoc representation of this doubly stochastic process.

While HMMs have already shown interesting performance in the context of CSR, there is still room for improving their quality. The reinforcement of discrimination between models appears one very promising issue. Until now, the main solutions suggested to compensate for the lack of discrimination in the HMM intervene in the training phase of the models (Bahl, Brown, de Souza, & Mercer, 1987; Ephraïm, Dembo, & Rabiner, 1989; Juang & Katagiri, 1992; Normandin, Cardin, & de Mori, 1994; Valtchev, Odell, Woodland, & Young, 1996). These techniques are rather complex and computationally expensive, although this later statement can be moderated as the extra cost only affects the training phase. Their major drawback lies in the fact that the statistics used by these methods cannot be computed exactly and then suboptimal estimates have to be used. An alternative consists in introducing a local discrimination ability in the model definition. The use of neural networks as a discriminant probability estimate has been shown to be relatively effective although expensive and both complex to implement and to tune (Bengio, 1993; Bourlard & Wellekens, 1989). Following the same basic idea, we propose to investigate a simple, effective and discriminant probability estimate: the k-nn estimate.

A major motivation for studying the k-nn estimate lies in its very low theoretical average classification error. It has also been shown that the gap between this error and the optimal Bayes error is directly linked to the value of k, assuming that a large enough training data set is available. This is an important asset compared with the continuous estimates such as Gaussian mixtures. The continuous estimates rely on the assumption that their elementary functions are consistent with the true data distribution. But the acoustic vectors generally do not follow Gaussian distributions (Montacie & Barras, 1996). And, even though it is theoretically possible to obtain a Gaussian mixture density which is arbitrarily close to the true distribution, the theory does not tell how this density should be constructed. For instance choosing the size of the mixture or training the weighting coefficients are actually guided by heuristics and the convergence towards the true distribution is not ensured. The non-parametric k-nn probability estimate offers an alternative to the continuous estimates which has never been studied in CSR on such a large scale. The main reason which has delayed the investigation of the k-nn estimate is its huge computational cost. This drawback has been addressed by the development of a fast algorithm which has allowed to reduce the k-nn cost to 0.2% of its initial value.

This study has been divided into three stages:

1. *Validation of the k-nn estimate local quality and comparison with the Gaussian mixture estimate.* In this first part, the k-nn estimate performance is evaluated on a local identification task of the TIMIT database;

2. *Introduction of the k-nn estimate into a baseline HMM system and comparative evaluation with a baseline Gauss/HMM system.* In this second step, the k-nn estimate is adapted to be introduced into the HMM formalism. To do so, the membership coefficients, measuring the degree of association between a vector and an HMM state, are introduced. A reestimation formula is presented and tested which adapts the Baum–Welch algorithm (Baum, 1972) to the training of the membership coefficients. From this point a baseline k-nn/HMM system has been defined based on an iterative EM training procedure whose convergence is proved. This system is compared with a baseline HMM system using Gaussian mixture state densities on a phone decoding of the TIMIT database;
3. *Improvements of the baseline system by integration of state-of-the-art techniques:* delta coefficients, gender modeling and contextual modeling with parameter tying are implemented and evaluated in the k-nn/HMM system.

The paper is organized as follows. In Section 2, the principle of the non-parametric probability estimation is presented along with some interesting properties of its average classification error rate. A baseline k-nn/HMM system is described in Section 3 and further refinements of the system are proposed in Section 4. Complete experimental setups and results are given in Section 5.

2. Non-parametric discriminant probability estimates

Non-parametric probability estimation makes no assumption of the targeted probability distribution shape. Therefore, no parameters are to be estimated. Their principle is a local estimation based on sample data.

Only the basic principles are introduced here; a more complete survey on the non-parametric estimation theory can be found in Fukunaga (1972) or Duda and Hart (1973).

2.1. Principle

Let $x_i, x_2, \ldots, x_n$ be n independent and identically distributed observations of a random variable X. The observations constitute the reference data set. The probability that a new observation x_0 will be located in a region R_n can be estimated by:

$$\hat{P}_n(x_0 \in R_n) = \frac{k_n}{n} \tag{1}$$

with k_n the number of observations of variable X included in the region R_n. $\hat{P}_n$ is an unbiased estimate of the true probability P. Therefore, an estimate by $\hat{p}_n$ of the probability density function p at x_0 can be expressed as:

$$\hat{p}_n(x_0) = \frac{\hat{P}_n(x_0)}{Vol_n} = \frac{k_n}{n \times Vol_n}, \tag{2}$$

where Vol_n is a measure of the volume of R_n (as a function of the amount of sample data n) which contains the k_n closest observations of x_0. It has been shown that $\hat{p}$ is an asymptotically unbiased and consistent estimate of p if Vol_n and k_n satisfy (the proof is omitted) (Loftsgaarden & Quesenbury, 1965):

$$\lim_{n\to\infty} Vol_n = 0, \tag{3}$$

$$\lim_{n\to\infty} k_n = \infty, \tag{4}$$

$$\lim_{n\to\infty} \frac{k_n}{n} = 0. \tag{5}$$

The first condition guarantees the convergence of $\hat{p}$ toward p by imposing the continuity of p in the neighborhood of x_0 (see mid-term in Eq. (2)). The second condition ensures the convergence in probability of k_n/n toward P (see Eq. (1)). Finally, the last condition implies that the ratio between the number of samples included in the region R_n and the total number of samples stays low, this is the condition to guarantee the convergence of $k_n/(n \times Vol_n)$ (see Fig. 1).

Values are to be chosen for the variables k_n and Vol_n. As these variables are dependent, only two methods can be considered:

- Setting a priori the regions R_n assuming that their respective volumes Vol_n conform to the previous conditions (3) and (5). For instance:

$$Vol_n = \frac{1}{\sqrt{n}}. \tag{6}$$

This method is known as the Parzen's kernel estimate (Parzen, 1962).

- Setting a priori the k_n as a function of n and increasing the volume Vol_n of R_n until k_n samples are included in R_n. This is the so-called *k nearest neighbors* estimate technique (Fix & Hodges, 1951).

2.2. Maximum a posteriori rule and classification error bounds

For classification purposes, the probability density values obtained from the k-nn estimate for each class are compared. The precision of the estimate can therefore be measured as its ability to correctly classify an observation according to the maximum a posteriori rule. It has been shown

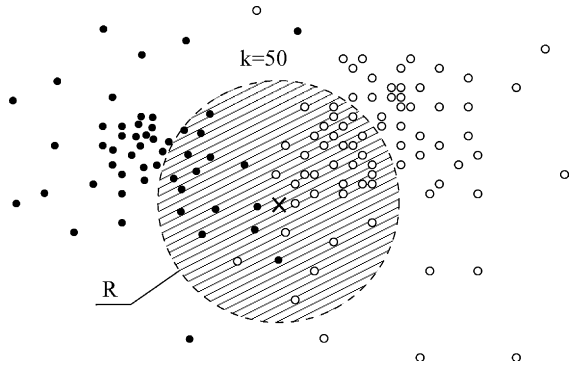

Fig. 1. Illustration of the k-nn principle in a two-dimensional representation space. Two classes are considered: white and black circles. In order to estimate the a posteriori probabilities of the vector $\times$, the region R has been enlarged so as to include the 50 nearest neighbors of $\times$. The estimates are $P(\times|\circ) = (35/50) = 0.7$ and $P(\times|\bullet) = (15/50) = 0.3$. So according to the maximum a posteriori classification rule, the vector $\times$ will be classified as a $\circ$.

that the k-nn estimate average classification error is always less than twice the optimal Bayes error. In addition, the error probability decreases monotonically when k increases (see for instance Fukunaga, 1972 for a proof of this propriety). An upper bound of the k-nn estimate error ε_k has been proposed as a function of k and the optimal error ε^* (Devijver & Kittler, 1982):

$$\varepsilon_k \leqslant \left(1 + \frac{2}{\sqrt{\pi \frac{k}{2}}}\right)\varepsilon^* \tag{7}$$

so, for example, in the 50-nn estimate case:

$$\varepsilon_{50} \leqslant 1.2\varepsilon^*. \tag{8}$$

Thus, the 50-nn estimate error lies between 100% and 120% of the Bayes error.

In summary, the decision rule derived from the k-nn estimate has a classification error close to the Bayes optimum which can be reduced by increasing k. This property constitutes one of the major assets of the method compared to the Gaussian mixture estimate: there exists a parameter which monotonically decreases the classification error. In the case of the Gaussian mixture estimate, the classification error reduction is generally obtained by increasing the number of Gaussians per mixture. This property, however, relies on the training techniques implemented and performance cannot be guaranteed.

3. *k*-nn/HMM system training

A new formulation of the state output probabilities needs to be derived in order to introduce the k-nn estimation into the HMM. This requires the association of the reference vectors to the HMM states.

A new formal framework is presented in which the degree of association between a reference vector and a state is quantified through a membership coefficient learnt on the training acoustic data. The Baum–Welch algorithm, widely used for the Gauss/HMM system parameters training, is adapted for the training of the membership coefficients.

3.1. Output state probability computation

In the HMM framework, since vectors are emitted by the model states, the reference vectors should be associated with the states. The output state probability for state i and vector x_0 given by the k-nn estimation is:

$$b_i(x_0) = \frac{k_i}{n_i}, \tag{9}$$

where k_i represents the number of nearest neighbors of x_0 out of k which are associated with the state i and n_i is the total number of reference vectors associated with state i.

By means of a manual phonetic transcription of the training data, a vector is first associated with the HMM representing its phonetic class. Then, the association between the vector and one of the HMM states is derived automatically. In this condition, the assignment of a

vector to a single state (*hard decision*) does not appear a desirable solution since it is obtained automatically.

The former consideration points out a general limitation of the k-nn estimate beyond the context of its introduction into an HMM system. The univocal association of a vector to a class does not offer a simple way to manage uncertainty on the reference data. This point can appear particularly delicate in domains where even the opinion of the human expert, when available, is not definitive. Some reference samples can be very characteristic of their class while others can be atypical.

Fuzzy set theory (Zadeh, 1965) can help to overcome this problem through the membership coefficients concept. Each reference vector x_0 is attributed a membership coefficient $u(x_0)$ for each possible class i such that:

$$u_i(x_0) \in [0,1], \tag{10}$$

$$\sum_{i=1}^{C} u_i(x_0) = 1 \tag{11}$$

with C the total number of classes. In our case, C is equal to S, the number of states in the system.

These coefficients can be set a priori by human experts in an arbitrary way or even evaluated (e.g., by combining the membership degrees of the point neighbors). An alternative even incorporates the distances to the neighbors in this combination (Keller, Gray, & Givens, 1985; Zouhal & Denoeux, 1997). On synthetic data, the derived decision rule has been shown to outperform the conventional k-nn rule (Zouhal & Denoeux, 1997).

This approach of a weighted association of the reference vectors to the classes has been formalized through the Dempster–Shafer theory (Denoeux, 1995). Beyond the considerations of recognition performance, this framework allows the introduction of a rich information into the recognition process, since it implicitly uses confidence measures. In the cited papers, this formalization leads to the improvement of the k-nn rule. However, in our case, the classification decision does not rely only on the k-nn probability estimate but on the HMM likelihood. Thus the membership coefficients are integrated in the expression of the HMM state output probability:

$$b_i(x_0) = \frac{\sum_{v=1}^{V} u_i(v) \times nn_k(x_0, v)}{U_i}, \tag{12}$$

where U_i represents the summation of the membership coefficients for state i over the V reference vectors and $nn_k(x_0, v)$ is an index function which is equal to 1 only if v is one of the k-nn of x_0:

$$nn_k(x_0, v) = \begin{cases} 1 & \text{if } v \in \{x : \ x \text{ is a } k\text{-nn of } x_0\}, \\ 0 & \text{otherwise.} \end{cases} \tag{13}$$

3.2. Training membership coefficients

The k-nn estimate parameters are the reference vector membership coefficients on the HMM states. The Baum–Welch algorithm being iterative, these parameters must be given initial values. The uniform segmentation of the training segments provides a first estimate of the membership

coefficients, adapted to the left-to-right HMM. The sequences are divided into as many contiguous subsequences as states in the HMM. Then, each vector of each subsequence is assigned to a state following the temporal order given by the transition parameters.

The reestimation formulae are expressed by means of the classical forward and backward variables, as defined in Baum (1972), for the transition parameters and the membership coefficients:

$$\hat{a}_{ij} = \frac{\sum_{t=1}^{T-1} \alpha_t(i) \times a_{ij} \times b_j(o_{t+1}) \times \beta_{t+1}(j)}{\sum_{t=1}^{T} \alpha_t(i) \times \beta_t(i)}, \tag{14}$$

$$\hat{u}_j(v) = \frac{\sum_{t=1}^{T-1} \sum_{i=1}^{N} (\alpha_t(i) \times a_{ij}) \times u_j(v) \times nn_k(o_{t+1}, v) \times \beta_{t+1}(j)}{\sum_{t=1}^{T-1} \sum_{i,j=1}^{N} (\alpha_t(i) \times a_{ij}) \times u_j(v) \times nn_k(o_{t+1}, v) \times \beta_{t+1}(j)}. \tag{15}$$

Convergence of the Baum–Welch algorithm using these formulae is shown in Appendix A. The reestimation formulae can be generalized to a set of examples by a separated summation over the numerator and the denominator:

$$\hat{a}_{ij} = \frac{\sum_{e=1}^{Ex} \sum_{t=1}^{T_e-1} \alpha_t^e(i) \times a_{ij} \times b_j(o_{t+1}^e) \times \beta_{t+1}^e(j)}{\sum_{e=1}^{Ex} \sum_{t=1}^{T_e} \alpha_t^e(i) \times \beta_t^e(j)}, \tag{16}$$

$$\hat{u}_j(v) = \frac{\sum_{e=1}^{Ex} \sum_{t=1}^{T_e-1} \sum_{i=1}^{N} (\alpha_t^e(i) \times a_{ij}) \times u_j(v) \times nn_k(o_{t+1}^e, v) \times \beta_{t+1}^e(j)}{\sum_{e=1}^{Ex} \sum_{t=1}^{T_e-1} \sum_{i,j=1}^{N} (\alpha_t^e(i) \times a_{ij}) \times u_j(v) \times nn_k(o_{t+1}^e, v) \times \beta_{t+1}^e(j)}. \tag{17}$$

Using these reestimation formulae, the convergence of the Baum–Welch algorithm is thus shown, at least towards the local maxima of the objective function. From this result, it has been possible to develop a baseline *k*-nn/HMM system with an optimal training protocol. The system performance is presented in Section 5.2.

4. Representation space and reference adaptations

In this section, ways to improve the baseline *k*-nn/HMM system are presented. They are largely guided by the state-of-the-art of the Gauss/HMM systems: enhancing the representation space by adding delta coefficients in the system features and adapting the references (gender and contextual modelings).

4.1. Integration of delta coefficients in the k-nn/HMM system

To combine the cepstral coefficients with the delta coefficients, a composite metric can be considered (Furui, 1986):

$$d_{\text{composite}} = d_M(\text{MFCC} + \text{E}) + w \times d_M(\Delta(\text{MFCC} + \text{E})), \tag{18}$$

where d_M is the Mahalanobis distance and w a weight reflecting the relative importance of the two subvectors MFCC + E and Δ(MFCC + E). With w set to 1, this technique is equivalent to concatenating the subvectors, a solution commonly used in the Gauss/HMM systems. In the case of

the k-nn/HMM system, out initial experiments have shown that w should be less than 0.05 to keep the k-nn computational cost reasonable. However, this would mean that the influence of the delta coefficients is nearly cancelled. It seems that the representation space combining static and dynamic dimensions has poor topological properties. This aspect does not have a large influence on the continuous estimates because of the assumption of dimension independence. A similar problem with the discrete estimate has been described in Lee (1988) where it was proposed to use multiple prototype dictionaries to overcome this shortcoming.

It seems preferable then to consider distinct representation spaces. The state output probability for the observation o_t is computed as the product of the output probabilities of several data streams (i.e., combination is made after the probability computation):

$$b_i(o_t) = \prod_F \left(b_i^f(o_t)\right)^{\gamma_f}, \tag{19}$$

where b_i^f is the emission probability of stream f given state i (applied to the corresponding subvector of o_t) and γ_f a weighting factor.

Data stream modeling is applied into the k-nn/HMM system by adapting the reestimation formulae (A.14) and (A.16). Each vector has two sets of membership coefficients: one associated with the static coefficient subvector $u_i^S(v)$ and the other with the delta coefficient subvector $u_i^\Delta(v)$. The membership coefficients corresponding to the static parameters stream (Eq. (15)) become:

$$\hat{u}_i^S(v) = \frac{\sum_{t=1}^{T}\sum_{j=1}^{N}[\alpha_{t-1}(j)\times a_{ji}]\times[u_i^S(v)]^{\gamma_S}\times nn_k^S(o_t,v)\times[b_i^\Delta(o_t)]^{\gamma_\Delta}\times\beta_t(i)}{\sum_{t=1}^{T}\sum_{j,i=1}^{N}[\alpha_{t-1}(j)\times a_{ji}]\times[u_i^S(v)]^{\gamma_S}\times nn_k^S(o_t,v)\times[b_i^\Delta(o_t)]^{\gamma_\Delta}\times\beta_t(i)}, \tag{20}$$

where

$$b_i^\Delta(o_t) = \frac{\sum_{v=1}^{V} u_i^\Delta(v)\times nn_k^\Delta(o_t,v)}{U_i}. \tag{21}$$

The S and Δ exponents imply that the function is applied to the corresponding static or dynamic subvector (MFCC + E or Δ(MFCC + E)). Weighting coefficients γ_S and γ_D are used to regulate the influence of each parameter group. The membership coefficient reestimation formula corresponding to the delta coefficients is obtained by exchanging the S and Δ indices in (20) and (21).

4.2. Gender modeling

A pattern recognition operator can be improved by appropriate modification of the a priori association between the sample data and the classes. In order to increase intra-class cohesion, it can be advantageous to divide the classes into subclasses. Applied to the phonetic classes, this principle leads to several types of modeling according to the subdivision criterion retained. One of the most straightforward of these criteria is the speaker gender.

Significant differences between male and female vocal tract characteristics having an incidence on acoustic realizations justify separate models for each gender. In the Gauss/HMM systems, gender modeling is generally obtained by a MAP adaptation of the global models. In the case of the k-nn/HMM system, the gender-dependent models are obtained by the partition of the reference set according to the speaker sex and the training of two model sets.

4.3. Contextual modeling

Another widely used subdivision criterion is co-articulation. Acoustic realizations of a phone are highly dependent on its phonetic context. This criterion leads to contextual modeling, which is investigated in the k-nn/HMM system. One major drawback of contextual modeling is that it greatly decreases the amount of data associated with each individual parameter. It is generally convenient to realize a tying of some parameters which means that they use common training data.

Parameter tying is applied to the context-dependent models at the state level. The Kullback–Leibler divergence (Kullback & Leibler, 1951), in its symmetrical version, is used to measure similarity between states. Applied between states i and j, the measure is:

$$KL(b_i||b_j) = \int_x b_i(x) \log \frac{b_i(x)}{b_j(x)} \, \mathrm{d}x + \int_x b_j(x) \log \frac{b_j(x)}{b_i(x)} \, \mathrm{d}x \tag{22}$$

with b_i the probability density function associated with state i.

Integration over the representation space can be transformed into a summation over the set V of all the reference vectors v:

$$\begin{aligned} KL(b_i||b_j) &= \sum_v \left(b_i(v) \log \frac{b_i(v)}{b_j(v)} - b_j(v) \log \frac{b_j(v)}{b_i(v)} \right) \\ &= \sum_v (b_i(v) - b_j(v)) \times (\log b_i(v) - \log b_j(v)). \end{aligned} \tag{23}$$

The divergence computation between two states requires the evaluation of the density functions on all the reference vectors. This procedure has a complexity in $o(VS^2)$ where V is the reference set size and S the total number of states in the system. With V being about 1 million and S around a few thousands, this procedure is very expensive. In order to obtain a reasonable cost, the number of vectors used for the divergence estimation has been reduced. A subset is randomly extracted from the reference set with a special care of maintaining the phonetic class proportions unchanged.

5. Experiments

To assess if it is worth integrating the k-nn estimate into the HMM formalism, we start by first evaluating the k-nn performance on a local identification task. Then, a baseline system is developed according to the above-defined protocol. Finally, integration of state-of-the-art techniques (delta coefficients, gender and contextual modeling) is considered.

5.1. Local identification with the k-nn rule

To carry out local identification experiments the implementation of the k-nn technique on a large-sized corpus requires first the development of a fast k-nn algorithm. Then the k-nn and the Gaussian mixture rules are evaluated and compared. Finally, the removal of the silence segments before the decoding is justified by measuring their influence on the k-nn rule.

5.1.1. Signal analysis and fast algorithm for k-nn computation

Numerous approaches have already been investigated in order to alleviate the huge computational cost of the k-nn estimate (Kim & Park, 1986; Miclet & Dabouz, 1983; Neimann & Goppert, 1988; Vidal, 1994). The algorithm used in this work, developed by C. Montacie and M.-J. Caraty, is based on the combination of three methods: Kittler–Richetin (Kittler, 1978), Friedman (Friedman, Bentley, & Finkel, 1975) and Fukunaga (Fukunaga & Narendra, 1975). The full description of this algorithm is not yet published but more details can be found in Lefevre, Montacie, and Caraty (1997) and Montacie, Caraty, and Lefevre (1997). This fast k-nn algorithm has been evaluated on TIMIT.

The DARPA TIMIT speech database (Garofolo et al., 1993) has been designed to provide acoustic-phonetic speech data for the development and evaluation of automatic speech recognition systems. It consists of utterances from 630 speakers representing the major dialects of American English. For each utterance, four elements are provided: the text, the signal sampled at 16 kHz (quantified on 16 bits) together with hand-labeled segmentation at the word level and at the phonetic level using 61 classes. Computed per centi-second, a vector is made of 12 MFCC and the short-term energy. The MFCC computation uses a 25 ms Hamming window and 24 intermediate spectrum filters. Subsequently, vectors are centered and normalized by the standard deviation evaluated on the training vector set. The training set is composed of 1,124,823 speech vectors (3696 sentences) and the core-test set [2] of 57,919 vectors (192 sentences). For each vector of the training and core-test set, the 50 nearest vectors are sought in the training set (which then becomes the *reference* vector set for the k-nn estimate). The gain in computation time appears to be around 99.8%. On a 600 MHz Pentium-PC, the 50-nn computation runs in 6 times real time, i.e., 16 50-nn are computed second.

5.1.2. k-nn/Gaussian mixture comparison on a local identification task

To carry out local identification of speech vectors, a decision rule is derived from the k-nn estimate. The decision rule according to the maximum a posteriori in a C-classes problem is:

$$D(x_0) = \text{argmax}_i(k_1, \ldots, k_i, \ldots, k_C), \tag{24}$$

where k_i is the number of nearest neighbors out of k associated with class i. This leads to a very simple decision rule: x_0 is affected to the class with the greatest k_i.

The local identification task principle is to assign each vector of the test set to the most probable phonetic class. We stress on the fact that this is accomplished without any segmental considerations (i.e., independently of the adjacent vectors). The k-nn rule is compared to a recognition operator based on the Gaussian mixture estimate. A Gaussian mixture is initialized for each class using the k-means algorithm (Linde, Buzo, & Gray, 1980) and its parameters are refined using the Baum–Welch algorithm.

Experiments are carried out on the 1,124,823 vectors of the TIMIT reference set (training set) and on the 57,919 vectors of the core test. The identification rate is obtained from a vector-by-vector comparison between the reference alignment provided with the database (Seneff & Zue, 1988) and the labeling proposed by the identification.

[2] The test data has a core portion containing 24 speakers (2 male and 1 female from each dialect region). This core-test, suggested by the TIMIT developers, generally produces results slightly worse than those of the whole test set.

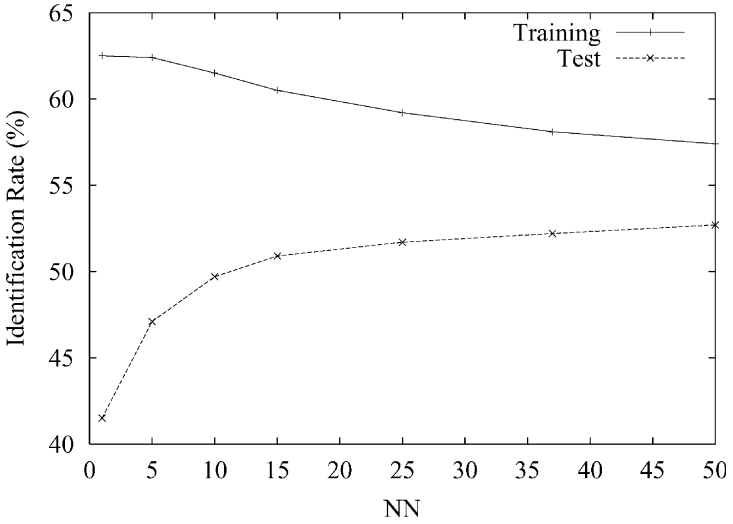

Fig. 2. Variation of the local identification rate for the k-nn decision rule as a function of k for the TIMIT training (plain line) and core-test (dashed line) sets.

Table 1
Local identification rate (%) of the k-nn decision rule as a function of k for the TIMIT training and core-test sets

nn	Training	Test
1	62.5	41.5
5	62.4	47.1
10	61.5	49.7
15	60.5	50.9
25	59.2	51.7
37	58.1	52.2
50	57.4	52.7

For the experiments on TIMIT, it is usual practice to authorize some confusions between some phones so that the 48 initial phonetic classes used during the decoding are brought back to 39 for the score computation (Lee, 1988). Some preliminary experiments have confirmed that this evaluation mode improves the Gaussian mixture estimate performance. However in the case of the k-nn estimate, using the 39 classes directly during the identification process leads to better results. For each estimate, the most favorable mode is used.

The results of the local identification task with the k-nn estimate, detailed in Table 1, are plotted in Fig. 2 for k varying from 1 to 50. In accordance with the theory presented in Section 2, the k-nn rule error decreases when k increases. A 11.2% identification gain is observed between the 1-nn and 50-nn rules. However, the identification rate of the training set falls by 5% over the same range. This phenomenon can be explained by the presence of vectors belonging to the same acoustic sequence (or at least to a sequence of the same speaker) among the *nn* of the training vectors. As k becomes larger, more vectors coming from different acoustic sequences are taken into account and the proximity effect tends to be reduced. This latter hypothesis seems to be confirmed by the common asymptote observed for the identification rate of the two evaluation

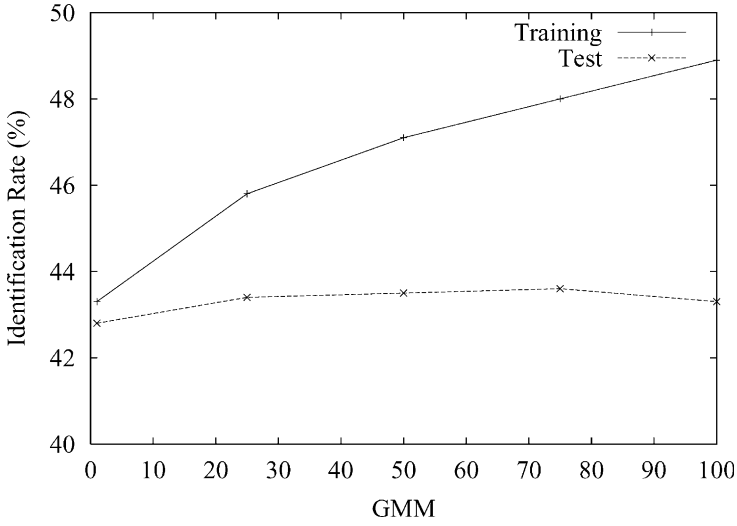

Fig. 3. Variation of the local identification rate for the Gaussian mixture-based decision rule as a function of the number of Gaussian functions per mixture for the TIMIT training (plain line) and core-test (dashed line) sets.

Table 2
Local identification rate (%) for the Gaussian mixture as a function of the number of Gaussian functions per mixture for the TIMIT training and core-test sets

No of Gaussian components	Training	Test
1	43.3	42.8
25	45.8	43.4
50	47.1	43.5
75	48.0	43.6
100	48.9	43.3

sets (around 55%, see Fig. 2). It can be concluded that increasing the number of *nn* improves the generalization ability of the *k*-nn estimate. Of course, this propriety should be observed only as long as the ratio k/n stays very low (see Eq. (5)).

Fig. 3 presents the results (detailed in Table 2) of the local identification task for the Gaussian mixture estimate. The number of Gaussian functions in the mixtures ranges from 1 to 100. The identification rate on the training set increases with the number of Gaussian functions per mixture (+5.2% from 1 to 100 functions). On the core-test set, the best score (43.6% correct identification rate) is obtained with a mixture of 75 Gaussian functions. This result suggests a specialization (*over-training*) of the Gaussian mixture estimate beyond a given number of Gaussian functions per mixture.

If the two best configurations for the *k*-nn and Gaussian mixture rules are considered for each set, the best identification rates on the training set are obtained with the 1-nn and 100-Gaussian mixture and on the test set, the 50-nn and 75-Gaussian mixture have led to the highest performances. In each case, the *k*-nn estimate achieves better results than the Gaussian mixture estimate (+9% on the test set and +14% on the training set). The scores being in the interval [40%,60%], the

Table 3
Local identification rates (%) for the 6 broad phonetic classes with the 50-nn and 75-Gaussian mixture estimates on the TIMIT core-test set

Broad phonetic class	50-nn	75-Gaussian mixture
Vowels	**39.7**	25.6
Liquids	35.9	**39.0**
Silence	**91.0**	76.3
Nasals	**42.6**	40.4
Fricatives	**52.1**	46.8
Stops	17.8	**25.8**

Best results in bold.

confidence interval radius for these experiments are approximately 0.1% for the training set and 0.4% for the core-test set.

Further experiments have been carried out to investigate the condensing (Hart, 1968) and editing (Wilson, 1972) techniques as a way to reduce the reference set size without degrading performance of the k-nn estimate. These experiments have shown that the reference set can be nearly halved without significant effect on the local identification result (Lefevre, Montacie, & Caraty, 1999).

5.1.3. Detailed analysis of the identification results

In order to evaluate more accurately the differences between the k-nn and the Gaussian mixture estimates, the local identification results for the 50-nn and 75-Gaussian mixture estimates are given in Table 3 on broad phonetic classes and on phone classes in Table 4.

The most important differences in performance between the 50-nn and the 75-Gaussian estimates are for silence (+14.7%) and vowels (+14.1%), Only the liquids (+3.1%) and the occlusives (+8%) obtain better results with the 75-Gaussian estimate. Detailed by phone classes, the differences are less clear: 24 classes out of 39 obtain better performance with the 75-Gaussian mixture estimate. These classes are characterized by a low number of samples in the test set. However, the class proportions in the test set follow those of the reference set. It appears that the k-nn estimate is sensitive to the amount of sample data per class. When this amount falls under a certain level (a few hundred vectors) the classes are almost completely misrecognized by the 50-nn estimate (e.g., /oy/, /uh/, /aw/, /y/, /ch/, /d/, /dx/ and /g/). This effect is compensated in the overall recognition rate by a good recognition of well-represented classes (e.g., /iy/, /ih/, /aa/, /er/, /l/, /n/, /f/, /s/ and /sil/). This phenomenon is particular to the k-nn probability estimation. With the Gaussian estimate, explicit models are associated to each class. During the identification process, only the model parameters are considered and thus the amount of sample data used during the training step has a less direct influence. In this latter regard, it can be concluded that the generalization ability of the k-nn estimation can be affected by sparsity in the representation space.

5.1.4. Speech/silence separation

Scoring of the TIMIT test set recognition generally takes into account the silence segments (see for instance Gauvain, Lamel, Adda, & Adda-Decker, 1994; Lee, 1988; Robinson, 1994). These segments represent 10% of the reference set (150k vectors out of 1.1 M). Within the framework of a non-parametric probability estimate, the presence of a dominant class is not desirable. For the

Table 4
Local identification rates (%) by phone class for the 50-nn and 75-Gaussian mixture estimates on the TIMIT core-test set

Phn	Vectors	50-nn	75-Gauss
Vowels			
iy	2127	**62.6**	32.0
ih	3461	**56.2**	20.4
eh	1696	**18.5**	17.2
ae	1458	**37.0**	31.1
aa	2898	**57.0**	36.4
ah	2016	**29.0**	19.5
uw	679	14.7	**23.4**
uh	220	0.0	**14.1**
er	2298	**59.0**	38.9
ey	1554	27.3	**27.8**
ay	1362	13.8	**14.8**
oy	294	0.0	**5.8**
aw	524	1.0	**12.8**
ow	1163	**17.1**	16.9
Liquids			
l	1941	**53.1**	43.2
r	1552	21.2	**31.6**
y	294	1.4	**35.7**
w	880	35.6	**44.1**
Silences			
sil	14,035	**91.0**	76.3
Nasals			
m	1274	36.9	**42.4**
n	2016	**51.8**	40.9
ng	312	6.7	**29.8**
Fricatives			
z	1530	23.1	**40.5**
v	553	22.1	**33.6**
f	1371	**64.0**	53.4
th	294	0.3	**32.0**
s	3661	**84.7**	55.4
sh	918	**67.5**	54.8
hh	519	12.3	**41.6**
ch	359	1.9	**26.5**
jh	274	5.1	**25.5**
dh	467	5.1	**22.7**
Stops			
b	228	2.6	**20.2**
d	292	1.4	**17.1**
dx	247	1.2	**20.6**
g	192	1.0	**30.2**
p	602	16.8	**34.6**
t	835	21.6	**29.6**
k	1523	**26.5**	23.2

Best results in bold. TIMIT alphabet is used for phone names.

Table 5
Local identification rate (%) of the 75-Gaussian mixture and 50-nn estimates after a speech/silence separation

Estimate	Training	Test
Standard		
75-Gaussian	48.0	43.6
50-nn	57.4	52.6
Silence decoded but not scored		
75-Gaussian	39.1	33.1
50-nn	46.9	40.4
Ideal speech/silence separation		
75-Gaussian	42.9	36.7
50-nn	52.1	45.8

k-nn estimate in particular, it is preferable that the class prior probabilities in the reference set conform to their real prior probabilities (Davies, 1988). And, as the silence is not part of the speech stream (exception made of short pauses between words), its prior can be fixed arbitrarily and then must not be compared with phone priors.

To avoid this shortcoming, the segments of silence at the beginning and end of the utterances are removed based on the reference data transcription. While automatic speech/silence segmentation algorithms would perform quite well in these conditions, an ideal separation has been retained in this study. The stop occlusions (voiced and unvoiced), previously included in the silence class, are now grouped in a stand-alone class. There are still 39 final phone classes but 46 intermediate classes (for the Gaussian mixtures), i.e., the 48 classes without /epi/ and /sil/. The results of the local identification, after discarding the segments of silence, are given in Table 5.

As expected, excluding silence substantially decreases the local identification rate since the silence class was correctly recognized and had a large amount of representative vectors. The identification rate when the silence segments are decoded but not scored falls to 40.4% for the 50-nn estimate (12.2% absolute loss). However, the identification rate obtained after the speech/silence separation is 45.8%. The separation allows a 5.4% gain (45.8–40.4%) on the identification of the other classes, due to the removal of confusions involving the silence class.

Latter on, an ideal a priori speech/silence separation will always be considered in the experiments.

5.2. Evaluation of baseline k-nn/HMM and Gauss/HMM systems

The *k*-nn/HMM system is evaluated and compared to a Gauss/HMM system on a phone decoding task. The baseline Gauss/HMM system used in this study is based on context-independent phone models, with a 3-state left-to-right HMM topology. The feature vectors contain 12 MFCC and the short-term energy (see Section 5.1.1). Each model is first trained using the Baum–Welch algorithm on each phone model and then refined by three iterations of connected reestimation. The decoding step is carried out by the Viterbi algorithm applied to a linear phone loop grammar. The language model is a phonetic back-off bigram with a unity cutoff. Previous experiments on TIMIT have shown that 32 Gaussians per mixture can be a reasonable balance between

Table 6
Recognition results (%) for the 32-Gauss/HMM and 50-nn/HMM systems on a phone decoding of TIMIT

Estimate	Corr	Sub	Del	Ins	Acc
Training					
32-Gaussian	59.7	21.7	18.6	1.8	57.9
50-nn	70.3	17.0	12.7	2.6	67.7
Test					
32-Gaussian	54.6	25.3	20.1	2.3	52.3
50-nn	57.6	27.8	14.6	5.0	52.5

performance and computational cost for phone decoding (Barras, 1996). Basically, the same configuration has been used for the 50-nn/HMM system with a few differences, such as the connected reestimation refinement, which has not been found profitable for the *k*-nn/HMM system. [3]

Comparative results using the 50-nn/HMM and the 32-Gauss/HMM systems are given in Table 6. Following the trend of the local identification results, a 10% increase is observed on the training data with the 50-nn system. But the performance of both systems is comparable on the core test (52.3% for the 32-Gauss/HMM system and 52.5% for the 50-nn/HMM system). However, the results of the systems are currently far from the best reported on TIMIT. State-of-the-art techniques are now considered in the *k*-nn/HMM system in order to improve its performance.

5.3. Integrating state-of-the-art techniques into the k-nn/HMM system

In the previous section the *k*-nn/HMM system has been shown to perform as well as a Gauss/HMM system in baseline conditions. The integration of state-of-the-art techniques into the *k*-nn/HMM system is now considered. Three techniques are investigated: delta coefficients, gender modeling and contextual modeling.

5.3.1. Delta coefficients

Two experiment sets are carried out to measure the delta coefficient contribution to the *k*-nn estimate: local identification and phone decoding. Each experiment is carried out with (Δ(MFCC + E)) and without delta coefficients (MFCC + E) in the system features. Delta coefficients of a particular vector are computed as the linear regression coefficients over a 5 vector centered window.

The local identification results are presented in Table 7. A significant improvement of the 50-nn performance is observed with the delta coefficients (around 3% on both sets). The improvement on the test set is larger with the 100-Gaussian estimate (8.6%) but its performance stays inferior to those of the 50-nn (45.3% vs. 49%).

A very significant increase in performance is generally observed with the introduction of the delta coefficients into the Gauss/HMM system: an 11.6% gain has been obtained on the training and test sets with our 32-Gauss/HMM system (Barras, 1996). However, the introduction of the

[3] Other slight differences, mainly the initialization procedure and the null probability treatment, are discussed thoroughly in Lefevre (2000).

Table 7
Correct identification rate (%) obtained with and without delta features of the 50-nn and 100-Gaussian mixture estimates

Estimate	Features	Training	Test
50-nn	MFCC + E	52.1	45.8
50-nn	MFCC + E + Δ(MFCC + E)	55.2	49.0
100-Gauss	MFCC + E	43.9	36.7
100-Gauss	MFCC + E + Δ(MFCC + E)	52.6	45.3

delta coefficients within the 50-nn/HMM system leads to a degradation of the performance: 6% on the training set and 2.3% on the test set as shown in Table 8.

In Section 5.1.2, it has already been observed that a performance improvement on the local identification task does not necessarily imply an improvement of the phone decoding performance. The results using the delta coefficients show that phone decoding can even be degraded while a gain is observed on the local identification task. We have actually no convincing explanation for this quite disappointing result although some hypotheses have already been investigated (Lefevre, Montacie, & Caraty, 1998).

5.3.2. Gender-dependent modeling

The gender is now used to split the training set in order to build gender-dependent models. As shown in Table 9, the gender modeling leads to a slight improvement of the global accuracy rate. Despite an 1.7% error rate reduction on the test set for the women, the overall gain is only 0.7% since there are more men than women in the core test.

Table 8
Recognition results (%) for the 50-nn/HMM system with and without delta features

Features	Corr	Sub	Del	Ins	Acc
Training					
MFCC + E	70.3	17.0	12.7	2.6	67.7
MFCC + E + Δ(MFCC + E)	70.6	19.2	10.2	8.8	61.8
Test					
MFCC + E	57.6	27.8	14.6	5.0	52.5
MFCC + E + Δ(MFCC + E)	61.8	27.5	10.7	11.6	50.2

Table 9
TIMIT phone decoding rates (%) obtained with gender-independent (GI) and gender-dependent (GD) 50-nn/HMM systems

Modeling	Corr	Sub	Del	Ins	Acc
Training					
GI	70.3	17.0	12.7	2.6	67.7
GD	70.8	16.6	12.6	2.5	68.3
Test					
GI	57.6	27.8	14.6	5.0	52.5
GD	57.9	27.0	15.2	4.7	53.2

5.3.3. Contextual modeling

Due to the size of the TIMIT database, the contextual modeling is limited to a right biphone modeling according to a previous study on the Gauss/HMM system (Barras, 1996). Right biphones occurring at least 50 times in the training set are retained. Considering 39 phone classes, 519 context-dependent models are obtained (out of the 1521 possible models): 480 right biphones to which are added the basic 39 monophones so as to model unrepresented biphones.

The state tying procedure, described in Section 4.3, is applied to the context-dependent models. The procedure is stopped when 200 states have been tied. Moreover, in order to reduce the Kullback–Leibler measure computation cost, only 10k reference vectors have been considered out of 1 million.

Phone decoding results are presented in Table 10. For each model set, the number of independent states is given. It appears that right biphone modeling (1557 states for 519 models) is not adapted to the k-nn/HMM system as it leads to a serious performance degradation (−9.3% on the training set and −20.9% on the test set). These results can largely be explained by a high level of insertion in the system. Some attempts to reduce the insertion rate by optimizing the value of the word penalty during the decoding have not led to better scores.

The results show that state tying has nearly no influence on the phone decoding accuracy for the context-dependent 50-nn/HMM system (see Table 10). Two hypotheses can explain this result: (1) the tying criterion is badly estimated because of too little data used for the Kullback–Leibler distance computation or (2) the number of tied states is still too low for a reliable estimation of the state probability densities. In any case, more fundamental objections can be raised to the use of context modeling in combination with the k-nn estimate. They are discussed in the next section.

5.4. Discussion

The experiments described in this section show that the techniques successfully applied to the Gauss/HMM systems are not straightforwardly applicable to the k-nn/HMM system. Regarding contextual modeling, some theoretical considerations can question its appropriateness in the k-nn/HMM framework.

Table 10
TIMIT phone decoding rates (%) for the 50-nn/HMM system with context-independent (CI), right biphone context-dependent (CD) and right biphone state-tied context-dependent (ST-CD) models with the total number of independent states in the models

Context	State No.	Corr	Sub	Del	Ins	Acc
Training						
CI	117 (39 × 3)	70.3	17.0	12.7	2.6	67.7
CD	1557 (519 × 3)	75.2	19.5	5.4	16.7	58.5
ST-CD	1357	74.8	19.7	5.5	16.6	58.2
Test						
CI	117 (39 × 3)	57.6	27.8	14.6	5.0	52.5
CD	1557 (519 × 3)	60.5	33.7	5.8	28.9	31.6
ST-CD	1357	60.7	33.6	5.7	29.3	31.5

To a certain extent, contextual modeling can be considered a solution, in the context of speech data, to the problem of mixture estimation. When a probability density function is estimated by mixtures of elementary functions (such as Gaussian), the data are divided into homogeneous subsets from which the individual function parameters are computed. If no particular criterion is available, the subsets are obtained in an unsupervised way, for instance with a k-means-like algorithm. In the case of the contextual modeling, an expert criterion is introduced based on the influence of the co-articulation at the phone realization level. The improvement of the estimate quality comes from a better construction of the data subsets associated to each function of the mixture. With regard to the k-nn technique, the estimation is local and does not rely on the accuracy of an a priori data partitioning. In light of these considerations, it seems that the contribution of the contextual modeling to the k-nn/HMM system could only be marginal.

On the other hand, the failure of the temporal parameters in the k-nn/HMM system can hardly be explained, particularly since an improvement is observed for local identification. A further study has been carried out in order to better understand the delta coefficient behavior in the Gauss/HMM system (Lefevre et al., 1998). However none of the various investigated hypotheses have allowed the derivation of a suitable technique for the introduction of the delta coefficients into the k-nn/HMM system.

6. Conclusion

The work presented in this paper is an attempt to increase the discrimination ability of the HMMs. In this goal, the k-nn probability estimate has been introduced for the first time into the HMM framework. Discriminant HMMs are usually obtained by further training iterations of the initial models based on a discriminative criterion which operates at the model level. The k-nn estimate gives us an opportunity to introduce discrimination at the HMM state level.

The development of a fast computation algorithm has drastically reduced the k-nn estimate cost and, as a consequence, has allowed its application to large amounts of speech data. In the experiments on the TIMIT speech database, the reference set includes more than one million acoustic vectors. In a first step, expectations from the theoretical properties of the k-nn estimate have been confirmed through local acoustic vector identification experiments. Under baseline conditions, the k-nn estimation has been shown to outperform the probability estimation based on mixtures of Gaussian. Thereafter, a first k-nn/HMM speech recognition system has been developed. To do so, an EM convergent training procedure based on the maximum likelihood criterion has been used. The trained parameters are the membership coefficients associating each reference vector to the HMM states.

A comparative evaluation of the baseline k-nn/HMM and Gauss/HMM systems on a phone decoding task has shown that their performances are of the same order. Several approaches were then investigated to improve the k-nn/HMM system performance: addition of delta coefficients into the acoustic features, gender modeling and contextual modeling with state tying. Other than a slight gain with gender modeling, none of these methods has improved the recognition results.

In summary, this work has followed a conventional framework to build a HMM-based system with a new state output probability estimation. It appears from the results that improvement of

the k-nn/HMM system will not be obtained from a straightforward application of the techniques used in the state-of-the-art HMM-based speech recognition systems. It is likely that this framework is really tied to the Gaussian estimation and does not generalize so well. To advance further, a framework more specific to the k-nn probability estimate should be proposed and evaluated.

In this idea, the study of the delta coefficients should be pursued as the inadequacy between the delta coefficients and the k-nn estimate remains an unexpected and unexplained result hindering the introduction of explicit temporal information into the k-nn/HMM system. In addition, numerous alternate ways remain open to improve the k-nn/HMM system performance. In particular, an effort is to be made on the development of topology inference techniques. The rigid model topology (a 3-state left-to-right topology) is potentially the reason why the improvement observed at the local level using the k-nn estimate is not reflected at the segmental level. The recourse to techniques in the field of grammatical inference (Miclet, 1990) should allow the development of a non-parametric approach of topology inference, which would be better suited to the k-nn probability estimation.

Appendix A. Proof of the reestimation formulae for the k-nn/HMM system parameters

The likelihood of a T-length vector sequence given a HMM is based on the integration of a hidden variable; the sequence of emitted states:

$$Q = q_1, \ldots, q_t, \ldots, q_T. \tag{A.1}$$

The sequence of nearest neighbors encountered during the emission is introduced as a second hidden variable:

$$P = p_1, \ldots, p_t, \ldots, p_T. \tag{A.2}$$

With N state HMMs, there exist N^T possible state sequences Q and k^T *nn* sequences P. The probability of a particular state and *nn* sequence is given by:

$$P(O, Q, P|\lambda) = a_{1_{q_1}} \prod_{t=1}^{T} b_{q_t}^{p_t}(o_t) a_{q_t q_{t+1}}, \tag{A.3}$$

where

$$b_{q_t}^{p_t}(o_t) = \frac{u_{q_t}(p_t) \times nn_k(o_t, p_t)}{U_{q_t}}, \tag{A.4}$$

where λ represents the HMM and $nn_k(o_t, p_t)$ is defined by (12). Eq. (A.4) represents the emission probability for a particular vector. An interpretation of (A.3) is that there exist N^T state sequences which could have generated the observation O and for each of theses sequences V^T branches to the reference vectors. In practice, only k^T branches are effectively encountered: the nn_k, function cancels the others. The total likelihood is obtained by summing Eq. (A.3) overall the possible P and Q sequences

$$P(O|\lambda) = \sum_{Q,P} \left\{ \prod_{t=1}^{T} b_{q_t}^{p_t}(o_t) \times a_{q_t q_{t+1}} \right\}. \tag{A.5}$$

The maximization of the HMM emission probability is obtained through an auxiliary function defined on the Kullback–Leibler measure (Kullback & Leibler, 1951):

$$F(\lambda,\hat{\lambda}) = \sum_{Q,P} P(O,Q,P|\lambda) \times \log P(O,Q,P|\hat{\lambda}) \tag{A.6}$$

with $\hat{\lambda}$ representing the HMM after reestimation. In Liporace (1982), it is shown that increasing F as a function of $\hat{\lambda}$ increases the likelihood $P(O|\hat{\lambda})$ of the observations given the model. The auxiliary function maximization can be divided into two distinct maximizations due to the logarithm:

$$\log P(O,Q,P|\hat{\lambda}) = \sum_{t=1}^{T} \log \hat{a}_{q_t q_{t+1}} + \sum_{t=1}^{T} \log \frac{\hat{u}_{q_t}(p_t)}{U_{q_t}} \times nn_k(o_t,p_t). \tag{A.7}$$

The nn_k index function is omitted in the log as it only results in a cancellation of the sequences not encountered and thus does not interfere with the magnitude of the log probability. The auxiliary function F can then be expressed as:

$$F(\lambda,\hat{\lambda}) = \sum_{i=2}^{N} F_{\hat{a}_t}(\lambda, \{\hat{a}_{ij}\}_{j=2}^{N} + \sum_{i=1}^{N} F_{\hat{u}_i}(\lambda, \{\hat{u}_i(v)\}_{v=1}^{V}), \tag{A.8}$$

where

$$\begin{aligned} F_{\hat{a}_t}(\lambda, \{\hat{a}_{ij}\}_{j=2}^{N}) &= \sum_{Q,P} P(O,Q,P,|\lambda) \sum_{t=1}^{T} \log \hat{a}_{q_t q_{t+1}} \times \delta_i^{q_t} \\ &= \sum_{j=2}^{N} \sum_{t=1}^{T} P(O, q_t = i, q_{t+1} = j|\lambda) \times \log \hat{a}_{ij}, \end{aligned} \tag{A.9}$$

$$\begin{aligned} F_{\hat{u}_i}(\lambda, \{\hat{u}_i(v)\}_{v=1}^{V}) &= \sum_{Q,P} P(O,Q,P|\lambda) \sum_{t=1}^{T} nn_k(o_t,p_t) \times \delta_i^{q_t} \times \log \frac{\hat{u}_{q_t}(p_t)}{U_{q_t}} \\ &= \sum_{v=1}^{V} \sum_{t=1}^{T} P(O, p_t = v, q_t = i|\lambda) \times nn_k(o_t,v) \times \log \frac{\hat{u}_i(v)}{U_i}, \end{aligned} \tag{A.10}$$

where δ is the Kronecker symbol, which is equal to 1 when both indices coincide, 0 otherwise. The maximization of $F_{\hat{a}_t}$ and $F_{\hat{u}_i}$ under the constraints:

$$\sum_{j=1}^{N} \hat{a}_{ij} = 1 \text{ and } \hat{a}_{ij} \geqslant 0 \text{ for every } i, \tag{A.11}$$

$$\sum_{v=1}^{V} \frac{\hat{u}_i(v)}{U_i} = 1 \text{ and } \hat{u}_i(v) \geqslant 0 \text{ for every } v \tag{A.12}$$

has the form $\sum_{i=1}^{I} x_i \log y_i$ with the constraints $\sum_{i=1}^{I} y_i = 1$ and $y_i \geqslant 0$. This expression reaches its global maximum for:

$$y_i = \frac{x_i}{\sum_{i=1}^{I} x_i} \text{ for every } i = 1 \text{ to } I. \tag{A.13}$$

The proof of this result can be found in Baum (1972). Intuitively, given the x_i, the total mass of the y_i (in fact 1) is divided so as to associate the highest y_i values to the highest x_i values. The following reestimation formula are then derived:

$$\hat{a}_{ij} = \frac{\sum_{t=1}^{T-1} P(O, q_t = i, q_{t+1} = j|\lambda)}{\sum_{t=1}^{T} P(O, q_t = i|\lambda)}, \tag{A.14}$$

$$\frac{\hat{u}_i(v)}{U_i} = \frac{\sum_{t=1}^{T} P(O, p_t = v, q_t = i|\lambda) \times nn_k(o_t, v)}{\sum_{t=1}^{T} P(O, q_t = i|\lambda) \times nn_k(o_t, v)}. \tag{A.15}$$

A solution in $\hat{u}_i$ to Eq. (A.15) under the condition (A.12) is given by:

$$\hat{u}_i(v) = \frac{\sum_{t=1}^{T} P(O, p_t = v, q_t = i|\lambda) \times nn_k(o_t, v)}{\sum_{t=1}^{T} P(O, p_t = v|\lambda) \times nn_k(o_t, v)} \tag{A.16}$$

under the assumption that the denominator of $\hat{u}_i$ has little variation as a function of v. Using this hypothesis, the probability that a reference vector belongs to the *nn* vectors of a particular segment is considered uniform.

Eqs. (A.14) and (A.16) can be evaluated by means of the classical forward/backward variables as defined in Baum (1972):

$$\hat{a}_{ij} = \frac{\sum_{t=1}^{T-1} \alpha_t(i) \times a_{ij} \times b_j(o_{t+1}) \times \beta_{t+1}(j)}{\sum_{t=1}^{T} \alpha_t(i) \times \beta_t(i)}, \tag{A.17}$$

$$\hat{u}_j(v) = \frac{\sum_{t=1}^{T-1} \sum_{i=1}^{N} (\alpha_t(i) \times a_{ij}) \times u_j(v) \times nn_k(o_{t+1}, v) \times \beta_{t+1}(j)}{\sum_{t=1}^{T-1} \sum_{i,j=1}^{N} (\alpha_t(i) \times a_{ij}) \times u_j(v) \times nn_k(o_{t+1}, v) \times \beta_{t+1}(j)}. \tag{A.18}$$

The reestimation formulae can be generalized to a set of examples by a separated summation over the numerator and the denominator:

$$\hat{a}_{ij} = \frac{\sum_{e=1}^{Ex} \sum_{t=1}^{T_e-1} \alpha_t^e(i) \times a_{ij} \times b_j(o_{t+1}^e) \times \beta_{t+1}^e(j)}{\sum_{e=1}^{Ex} \sum_{t=1}^{T_e} \alpha_t^e(i) \times \beta_t^e(j)}, \tag{A.19}$$

$$\hat{u}_j(v) = \frac{\sum_{e=1}^{Ex} \sum_{t=1}^{T_e-1} \sum_{i=1}^{N} (\alpha_t^e(i) \times a_{ij}) \times u_j(v) \times nn_k(o_{t+1}^e, v) \times \beta_{t+1}^e(j)}{\sum_{e=1}^{Ex} \sum_{t=1}^{T_e-1} \sum_{i,j=1}^{N} (\alpha_t^e(i) \times a_{ij}) \times u_j(v) \times nn_k(o_{t+1}^e, v) \times \beta_{t+1}^e(j)}. \tag{A.20}$$

Using these reestimation formulae, the Baum–Welch algorithm is thus shown to converge, at least towards the local maxima of the auxiliary function.

References

Bahl, L., Brown, P., de Souza, P., Mercer, R., 1987. Speech recognition with continuous-parameter hidden Markov models. Computer Speech and Language 2, 219–234.

Barras, C., 1996. Reconnaissance de la parole continue: adaptation au locuteur et contrôle temporel dans les modèles de Markov cachés. Ph.D. thesis, Université Pierre et Marie Curie, Paris.
Baum, L., 1972. An inequality and association maximization technique in statistical estimation for probabilistic functions of Markov processes. Inequalities 3, 1–8.
Bengio, Y., 1993. A connectionist approach to speech recognition. International Journal of Pattern Recognition and Artificial Intelligence 7 (4), 647–667.
Bourlard, H., Wellekens, C., 1989. Speech pattern discrimination and multilayer perceptrons. Computer Speech and Language 3, 1–19.
Davies, E., 1988. Training sets and a priori probabilities with the nearest neighbour method of pattern recognition. Pattern Recognition Letters 8, 11–13.
Denoeux, T., 1995. A *k*-nearest neighbor classification rule based on Dempster–Shafer theory. IEEE Transactions on Systems, Man and Cybernetics 25 (5), 1–28.
Devijver, P., Kittler, J., 1982. Pattern Recognition: A Statistical Approach. Prentice-Hall, London.
Duda, R., Hart, P., 1973. Pattern Classification and Scene Analysis. Wiley, New York.
Ephraïm, Y., Dembo, A., Rabiner, L., 1989. A minimum discrimination information approach for hidden Markov modeling. IEEE Transactions on information Theory 35 (5), 1000–1013.
Fix, E., Hodges, J., 1951. Discriminatory analysis: nonparametric discrimination: consistency proprieties. Report 4, USAF School of Aviation Medicine, February 1951.
Friedman, J., Bentley, J., Finkel, R., 1975. An algorithm for finding nearest neighbors. IEEE Transactions on Computers, 1000–1006.
Fukunaga, K., 1972. Introduction to Statistical Pattern Recognition. Academic Press, New York.
Fukunaga, K., Narendra, P., 1975. A branch and bound algorithm for computing k-nearest neighbors. IEEE Transactions on Computers, 750–753.
Furui, S., 1986. Speaker-independent isolated word recognition using dynamic features of speech spectrum. IEEE Transactions on ASSP 34 (1), 52–60.
Garofolo, J.S., Lamel, L.F., Fisher, W.M., Fiscus, J.G., Pallett, D.S., Dahlgren, N.L., 1993. The DARPA TIMIT Acoustic-Phonetic Continuous Speech Corpus CDROM ntis order number pb91-100354 edition.
Gauvain, J.-L., Lamel, L., Adda, G., Adda-Decker, M., 1994. Speaker-independent continuous speech dictation. Speech Communication 15, 21–37.
Hart, P., 1968. The condensed nearest neighbor rule. IEEE Transactions on Information Theory 14, 515–516.
Juang, B.-H., Katagiri, S., 1992. Discriminative learning for minimum error classification. IEEE Transactions on Signal Processing 40 (12), 3043–3054.
Keller, J., Gray, M., Givens, J., 1985. A fuzzy *k*-nn neighbor algorithm. IEEE Transactions on Systems, Man and Cybernetics 15 (4), 580–585.
Kim, B., Park, S., 1986. A fast *k* nearest neighbor finding algorithm based on the ordered partition. IEEE Transactions on Pattern Analysis and Machine Intelligence 8 (6), 761–766.
Kittler, J., 1978. A method for determining *k*-nearest neighbours. Kibernetics 7, 313–315.
Kullback, S., Leibler, R., 1951. On information and sufficiency. Annals of Mathematical Statistics 22, 79–86.
Lee, K.-F., 1988. Large-vocabulary speaker-independent continuous speech recognition: the SPHINX system. Ph.D. thesis, Carnegie Mellon.
Lefevre, F., 2000. Estimation de probabilité non-paramétrique pour la reconnaisance markovienne de la parole. Ph.D. thesis, Université Pierre et Marie Curie, Paris.
Lefevre, F., Montacie, C., Caraty, M.-J., 1997. *K*-nn estimator in a HMM-based recognition system. In: Ponting, K. (Ed.), Computational Models of Speech Pattern Processing. NATO ASI-Series F. Springer, Berlin, pp. 96–101.
Lefevre, F., Montacie, C., Caraty, M.-J., 1998. On the influence of the delta coefficients on a HMM-based recognition system. In: Proceedings ICSLP/SST, Sydney.
Lefevre, F., Montacie, C., Caraty, M.-J., 1999. A mle algorithm for the *k*-nn/HMM system. In: Proceedings ESCA, Eurospeech, Budapest, vol. VI, pp. 2733–2736.
Linde, Y., Buzo, A., Gray, R., 1980. An algorithm for vector quantizer design. IEEE Transactions on Communications 28 (1), 84–95.

Liporace, L., 1982. Maximum likelihood estimation for multivariate observations of Markov sources. IEEE Transactions on Information Theory 28 (5), 729–734.

Loftsgaarden, D., Quesenbury, C., 1965. A nonparametric estimate of a multi-variate density function. Annals of Mathematical Statistics 36, 1049–1151.

Miclet, L., 1990. Grammatical inference. In: Bunke, H., Sanfeliu, A. (Eds.), Syntactic and Structural Pattern Recognition: Theory and Application. Series of Computer Science, vol. 7. World Scientific, Singapore, pp. 237–290.

Miclet, L., Dabouz, M., 1983. Approximative fast nearest-neighbor recognition. Pattern Recognition Letters 1, 277–285.

Montacie, C. Caraty, M.-J., Barras, C., 1996. Mixture splitting technic and temporal control in a HMM-based recognition system. In: Proceedings ICSLP, Philadelphia, vol. III, pp. 977–980.

Montacie, C., Caraty, M.-J., Lefevre, F., 1997. *K*-nn versus Gaussian in a HMM-based system. In: Proceedings ESCA Eurospeech, Rhodes, vol. II, pp. 529–533.

Neimann, H., Goppert, G., 1988. An efficient branch-and-bound nearest neighbour classifier. Pattern Recognition Letters 7, 67–72.

Normandin, Y., Cardin, R., de Mori, R., 1994. High-performance connected digit recognition using maximum mutual information estimation. IEEE Transactions on Speech and Audio Processing 2 (2), 299–311.

Parzen, E., 1962. On estimation of a probability density function and mode. Annals of Mathematical Statistics 33, 1065–1076.

Robinson, A., 1994. An application of reccurent nets to phone probability estimation. IEEE Transactions on Neural Networks 5 (2), 298–305.

Seneff, S., Zue, V., 1988. Transcription and Alignment of the TIMIT Database. Getting Started with the DARPA CD-ROM: An Acoustic-Phonetic Continuous Speech Database, NIST.

Valtchev, V., Odell, J., Woodland, P., Young, S., 1996. Lattice-based discriminative training for large vocabulary speech recognition. In: Proceedings IEEE ICASSP, Atlanta, vol. I, pp. 605–608.

Vidal, E., 1994. New formulation and improvements of the nn asea. Pattern Recognition Letters 15 (1), 1–7.

Wilson, D., 1972. Asymptotic proprieties of nearest neighbor rules using edited data. IEEE Transactions on Systems, Man and Cybernetics 2 (3), 408–421.

Zadeh, L., 1965. Fuzzy sets. Information Control 8, 338–353.

Zouhal, L., Denoeux, T., 1997. Generalizing the evidence-theoretic *k*-nn rule to fuzzy pattern recognition. In: Second International Symposium on Fuzzy Logic and Applications, Zurich, pp. 294–300.