On the Influence of the Delta Coefficients in a HMM-based Speech Recognition 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

The delta coefficients are a conventional method to include temporal information in the speech recognition systems. In particular, they are widely used in the gaussian HMM-based systems. Some attempts were made to introduce the delta coefficients in the K-Nearest Neighbours (K-NN) HMM-based system that we recently developed. An introduction of the delta coefficients directly in the representation space is shown not to be suitable with the K-NN probability density function (pdf) estimator. So, we investigate whether the delta coefficient could be used to improve the K-NN HMM-based system in other ways. In this purpose, an analysis of the delta coefficients in the gaussian HMM-based systems is proposed. It leads to the conclusion that the delta coefficients influence also the recognition process.

1. INTRODUCTION

We have recently proposed the use of the K-NN pdf estimator as an alternative to the widespread gaussian pdf estimator [1]. The K-NN estimator is known as one of the best probability estimator provided that enough training data are available. A huge computational effort, its main drawback, has been seriously reduced by the development of a fast K-NN algorithm. The estimator has been integrated with few peculiarities in the HMM framework. The first experiments on a continuous recognition task have shown comparable performance between an 8-gaussian HMM-based system and a K-NN HMM-based system [2].

Thereafter, we have intended to upgrade the K-NN HMM-based system with the state-of-the-art HMM techniques. The introduction of the delta coefficients in the representation space has seemed the most straightforward technique to implement. Since its introduction [3], this technique has been applied successfully to most speech recognition tasks. Other ways to incorporate temporal information in the features of a HMM-based system have been investigated, such as Spectral Variation Function [4], but none had allied so much effectiveness and simplicity than the delta coefficients. From a discriminative point of view, some studies have even shown that it was preferable to keep the couple of corresponding static and delta coefficients together rather than to enlarge the number of static coefficients in the representation space [5]. Thus, the delta coefficients were introduced in the K-NN HMM-based system.

In a first attempt, the static and delta coefficients are used jointly, which is the conventional implementation in the gaussian HMM-based systems. In a second attempt, the static and the delta coefficients are split into two streams. The total pdf was then obtained by multiplying the two stream pdfs. Both methods offered no improvement compared to the baseline system. These attempts show that the delta coefficients could not be implement in a K-NN HMM-based system as a simple growth of the representation space.

However, an evaluation –described in Section 2- shows that the improvement in a gaussian HMM-based system using delta coefficients is not only due to a better representation space but also to a better segmentation during the recognition process. This latter point is to be clarified for being reproduced in the K-NN HMM-based system. Two hypotheses are studied. The first one results from a side effect of the use of the gaussian pdf estimator: since they are not normalised, the gaussian pdfs could assign very high output probabilities to certain frames during the recognition process. These frames will act then as anchor points of the recognition process. This particular problem is shown to be related to the use of the delta coefficients. Finally, this point is shown -in Section 3- to have no influence on the recognition results. The second hypothesis -developed in Section 4- is that the delta coefficients induce a better fitting of the data to the HMM topology. An entropy measure confirms this assumption. Then a new method directly connected to the temporal information brings

by the delta coefficients is presented to integrate them with much more profit in the K-NN HMM-based system.

2. EVALUATION OF THE DELTA COEFFICIENTS

The improvement due to the introduction of the delta coefficients in a gaussian HMM-based system is well-known for continuous recognition tasks. But, the evaluation of the whole system also requires the evaluations of its different components so as to get insights in their contributions to the global improvement. The evaluated components are: the pdf estimator, the dynamic warping within the HMMs and the dynamic warping between the HMMs.

2.1. Evaluation Paradigm

The evaluation paradigm responds to the above-mentioned requirements through three series of experiments:

- *Frame identification*
  This experiment aims at assessing the pdf estimators apart from any segmental modelling. The principle of the frame identification is to assign each frame issued from the test data to its most probable phonetic class without any segmental considerations. The decision rule is straightforward with the K-NN estimator: the most probable class is the class having the highest number of nearest-neighbours of the frame. For the gaussian estimator, a gaussian mixture has to be learned for each class. The mixtures can be obtained using a k-means algorithm. Practically, the gaussian mixtures we use come from the training a gaussian HMM-based system. So, each class was not represented by a single mixture but by a set of N mixtures corresponding to the N states of its HMM. The class probability of the frame is the highest of its gaussian mixture values. The correct rate is computed from a one-to-one frame comparison between the reference labelling and the identification labelling.
- *Segmental identification*
  This experiment provides an assessment of the pdf estimators introduced in the HMM framework. In this task, the boundaries of the phonemes are supposed to be known. Each phonetic class is represented by a HMM. The gaussian pdf parameters are learned by the Baum-Welch algorithm. The test-segments are warped with all the models using a Viterbi algorithm. The identified class corresponds to the model with the highest warping likelihood.
- *Continuous recognition*
  This experiment is similar to the previous one except that the boundaries of the segments are unknown. The sentences are warped using the Token Passing algorithm [6] with a phonetic bigram language model. The accuracy rate takes account of the deletions and the insertions of phonemes.

Each experiment is performed with 3 different pdf estimators: the gaussian estimator with static coefficients, the gaussian estimator with static and delta coefficients and the 50-NN estimator. The gaussian pdf estimator is an 8-gaussian mixture. The HMMs are 3-states left-to-right Bakis model with central skip.

Computed each centi-second, a frame is represented by 12 Mel-Frequency Cepstrum Coefficients and by the energy coefficient. The delta coefficients are computed using the following regression formula:

$$D_i(t) = \frac{\sum_{n=1}^{+n0} n \cdot (S_i(t+n) - S_i(t-n))}{2\sum_{n=1}^{+n0} n^2}$$

where D_i is a delta coefficient at time t corresponding to the static coefficient S_i. These coefficients approximate the slope of the time variation of the corresponding static coefficients. The length of the interval was set to five frames ($n0$=2), representing 50 ms. The experiments were carried out on the TIMIT database. The core-test is the usual subset of the test (192 sentences including 57,919 frames). The TIMIT reference-labelling [7] is used for the classification of the training frames. Figure 1 summarises the results on the core-test for the three series of experiments. The results are detailed by phonemes in Table 2 (inserted at the end of the paper).

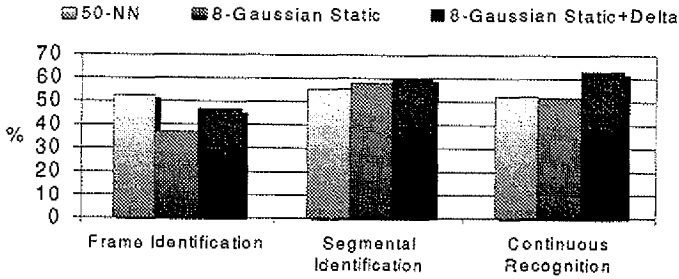

Figure 1: Results for the three series of experiments on the TIMIT core-test

2.2. Evaluation Analysis

As far as only pdf estimator is considered, it is noticeable that the 50-NN estimator is of better quality than the gaussian estimator. The frame identification reveals that the 50-NN estimator obtains better identification rate even when the gaussian estimator is used with twice as many coefficients. On the other hand, the segmental identification and continuous recognition show that this quality is not much reinforced by the introduction of the 50-NN estimator within the HMM framework.

Considering the gaussian estimator, the introduction of the delta coefficients increases the results in the three experiments. Nonetheless, these improvements are not equally shared among all the phonemes (cf. Table 2). But no particular phonetic macro-class seems to benefit more. Moreover, it can be stated that the gaussian HMM-based system used without delta coefficients has its performance decreasing between the segmental identification and the continuous recognition. A high level of deletions accounts for that fact. Whereas, with delta coefficients, the gaussian HMM-based system makes profit of the segmentation to correct ambiguous segments during the continuous recognition process.

From this last point, it can be concluded that the delta coefficients have a beneficial influence in the segmentation during the recognition process. In a fist step, we will verify that this does not result from a side effect of the gaussian pdf estimator computation.

3. ANCHOR POINTS DURING RECOGNITION PROCESS

In the theoretical formulation of the HMM, the states are emitting symbols with a given output probability. With non-parametric representations, these probabilities are directly estimated from real data points. With continuous parametric representations, since probability distributions are considered, the probabilities are extrapolated from probability density functions. One major difference is the loss of normalisation. Most of the time, since variances are very similar among all pdfs, the pdf values represent a convenient estimate of the output probabilities. But, when the variances become different and low-valued, some "resonance" effect can appear. The "resonance" effect is characterised by pdfs giving very high output probability values to their most probable frames. This frames can then be considered as anchor points of the recognition process.

In the case of the gaussian pdfs, the output probability of state s for the frame vector o_t of dimension n is computed from the formula:

$$b_s(o_t) = \sum_{m=1}^{M} c_{sm} \times \left(F \cdot \exp\left(-\frac{1}{2}(o_t - \mu_{sm})^T \Sigma_{sm}^{-1}(o_t - \mu_{sm})\right)\right) \text{ with } F = \frac{1}{\sqrt{(2\pi)^n |\Sigma_{sm}|}}$$

where M is the number of mixture components, c_{sm} is the weight of the m^{th} component, μ_{sm} is the mean vector and Σ_{sm} the covariance matrix. From this formula, it can be seen that F is the maximum probability value that can be reached by the gaussian pdf. Thus, when F becomes very high, frames closed to the mean vector will produce a "resonance" effect. Figure 2 represents the histogram of the

logarithms of factor F, computed from the 1,152 gaussian pdfs of the 8-gaussian HMM-based system with static coefficients and with static and delta coefficients. As the pdfs estimate probabilities, F being equal or inferior to 1 seems a reasonable constraint. If 90% of the gaussian pdfs could conform to this constraint when static coefficients are used, this proportion falls to 7% when delta coefficients are involved.

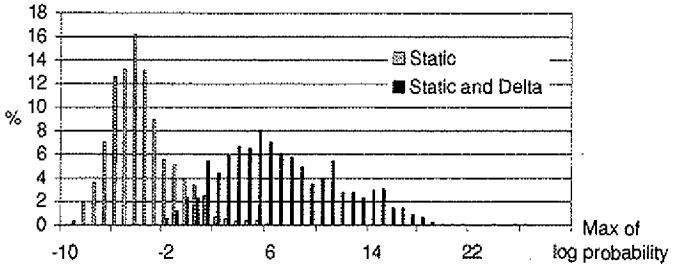

Figure 2: Histogram of the maximum of log probability in a 8-gaussian HMM-based system

The impact of the "resonance" effect is estimated on the continuous recognition task on the TIMIT core-test (cf. third experiment in previous section). The "resonance" effect is not marginal since it occurs with 23% of the 57,919 frames. Some attempts are made for removing the "resonance" effect. In the first case, the output probabilities overstepping 1 are cut to 1. Somehow crude, this method has only decreased the recognition rate by a 1%, which is under the confidence interval. In the second case, when a "resonance" effect appears, the output probabilities of all states for that frame were divided by the "resonance" peak so as to keep the better output probability always under or equal to 1. This experiment has shown no difference in the recognition rate.

These experiments do not show that the recognition process is sensitive to the "resonance" effect. Moreover, the silence classes (*cl, sil*) account for 30% of all the "resonance" frames and yet they were correctly recognised without delta coefficients (see Table 2). Then, the influence of the delta coefficients on the recognition process does not appear to be the fact of the pdf computation. Consequently, we look for the improvement due to the delta coefficients in the relation between the data and the HMM topology.

4. FITTING OF THE DATA TO THE HMM TOPOLOGY

If we refer to information theory, an HMM topology can be considered well fitted to the data when its states plays an almost equal role. This means, theoretically, that more the model is closed to the equiprobability more it will be efficient. The model entropy is a good criterion of the "goodness of fit".

The HMM is regarded as a source from which could arise L different symbols. In this case, the symbols are the states of the model. The computation of the model entropy requires the estimation of the frequencies of occurrence of the states. These frequencies are related to the unknown (*hidden*) state sequences of the data segments through their corresponding models. The segments are warped using the Viterbi algorithm, which provides the best state sequence. The model entropy can then be compared to the theoretical one corresponding to an equiprobable model. A source emitting L symbols with equal probability has an entropy value of $\log_2 L$. In the case of 3 states Bakis models, L equals 3. The average model entropy values are given in Table 1 for the 8-gaussian HMM-based system on the TIMIT core-test used with static coefficients (S) and with static plus delta coefficients (S+D).

Table 1 shows that the models used with delta coefficients have an entropy closer to the theoretical entropy of a 3-states model. This result confirms the relation between the introduction of the delta coefficients and a better fitting of the data to the HMM topology. To take advantage of these preliminary results, an attempt is made by restricting the use of the delta coefficients in a K-NN HMM-based system to their temporal aspect.

Systems	S	S+D	Theoretical Maximum
Entropy	1.51	1.56	1.59

Table 1: Average entropy for the systems s with static coefficients (S) and the models with static and delta coefficients (S+D)

In the purpose, the delta coefficients are no more considered to bring phonetic information and are labelled according to time classes (e.g., start, middle and end). The static and delta coefficients are split in two streams. The static stream gives a phonetic probability to each model and the delta stream gives a temporal probability to the states. The temporal probability due to the delta stream depends only on the order of the considered states (the first state is associated with the start time class, and so on). The state output probability is obtained by multiplying the two streams. The preliminary experiment of this new technique on a continuous recognition has not yielded to improvement.

5. CONCLUSION

Some attempts have been made to improve the K-NN HMM-based system by the introduction of the delta coefficients. Their simple addition to the representation space results in no improvement since this method is not suitable for the K-NN pdf estimator. Then, a new technique has to be found to benefit from the temporal information of the delta coefficients in the K-NN HMM-based system.

An analysis of the delta coefficients in the gaussian HMM-based system has shown that they have a beneficial influence in the recognition process. A first experiment is made to verify whether this influence is not due to a side effect of the gaussian pdf computation. As this hypothesis is not verified, we have investigated the relation between the introduction of the delta coefficients and a better fitting of the data to the HMM topology. This hypothesis has been verified through an entropy measurement.

A new method is presented for integrating the delta coefficients in the K-NN HMM-based system. The principle is to use the delta coefficients considering temporal classes rather than phonetic classes. No significant improvement has been obtained, but this method will be investigated more thoroughly.

Our future works will also consist in improving the introduction of the K-NN pdf estimator in the HMM framework since the K-NN HMM-based system does not benefit as it should from the quality of the K-NN pdf estimator. If we refer to the gaussian HMM-based system, a substantial improvement could be obtained.

6. REFERENCES

[1] Montacié, C., M.-J. Caraty and F. Lefèvre (1997). *KNN versus Gaussian in a HMM-Based System.* Eurospeech, Rhodes.

[2] Lefèvre F., C. Montacié and M.-J. Caraty (1997). *K-NN Estimator in a HMM-Based Recognition System* in Computational Models of Speech. NATO ASI, Jersey.

[3] Furui, S. (1986). Speaker-independent Isolated Word Recognition Using Dynamic Features of Speech Spectrum. IEEE Trans. on ASSP 34(1): 52-60.

[4] Brugnara, F., R. de Mori, D. Giuliani and M. Omologo. (1992). *Improved Connected Digit Recognition Using Spectral Variation Functions.* ICSLP, Banff.

[5] Bocchieri, E. and J. Wilpon (1993). *Discriminative Feature Selection for Speech Recognition.* Computer Speech and Language 7: 229-246.

[6] Young S.J., N.H. Russel and J.H.S. Thornton (1989). *Token Passing: a Simple Conceptual Model for Connected Speech Recognition Systems.* Technical Report, CUED-Speech Group, Cambridge.

[7] Seneff, S. and V. Zue (1988). Transcription and Alignment of the TIMIT Database, NIST.

	Frame Identification			Segmental Identification			Continuous Recognition		
Vowels									
iy	63.7	32.6	42.8	67.2	52.7	67.1	82.6	62.1	82.1
ih	57.5	13.6	19.2	67.5	43.1	45.3	73.6	61.0	68.5
eh	19.1	18.9	20.8	42.8	26.5	41.3	40.0	39.0	49.3
ae	35.3	29.5	34.8	35.6	41.0	44.8	62.2	53.1	60.2
ah	28.6	23.6	24.2	24.6	39.3	45.2	33.9	51.8	63.0
uw	15.6	24.6	33.6	14.9	48.6	54.1	25.9	61.1	65.1
uh	0.0	12.7	21.8	0.0	24.1	24.1	0.0	22.7	31.0
aa	56.3	19.7	28.5	58.8	50.0	68.4	71.4	61.1	75.2
er	57.5	39.0	48.5	59.7	51.7	63.7	76.7	63.8	74.6
ey	26.3	31.6	44.7	41.7	56.1	67.5	55.3	65.3	75.2
ay	13.6	15.2	36.0	51.7	47.2	66.3	50.6	55.6	68.3
oy	0.0	5.1	22.1	0.0	50.0	62.5	0.0	21.4	46.7
aw	0.6	17.8	36.1	0.0	33.3	60.0	4.2	43.5	63.3
ow	17.7	15.6	29.0	25.0	34.8	48.3	50.0	47.7	59.4
Liquids									
l	52.8	46.1	49.6	67.5	56.4	66.7	67.6	62.8	67.2
r	23.0	27.3	38.9	53.9	50.4	64.4	55.2	65.8	74.4
y	1.4	34.7	54.8	2.0	56.0	66.0	14.3	64.0	75.7
w	35.9	38.8	53.1	42.5	56.9	61.8	62.3	71.2	81.2
hh	11.6	34.5	49.5	11.1	61.0	31.2	23.9	63.8	82.5
Nasals									
m	36.7	36.1	48.6	47.9	47.8	47.8	43.9	61.4	64.5
n	51.8	17.7	28.3	55.6	49.6	56.6	79.9	59.3	70.2
ng	6.7	33.3	49.4	3.9	46.2	61.5	0.0	38.5	65.9
Stops									
b	2.2	16.7	52.6	8.5	44.7	41.7	30.9	66.7	73.0
d	1.4	18.8	33.2	9.6	38.4	18.8	24.3	49.2	57.1
dx	1.2	21.9	32.4	4.4	47.8	28.9	27.1	30.6	75.0
g	0.5	28.1	44.3	3.6	49.2	36.9	14.0	51.1	54.4
p	18.6	31.2	48.8	45.8	46.3	22.1	60.0	62.4	71.9
t	23.7	29.3	40.0	40.2	51.6	34.4	82.4	76.3	84.5
k	26.2	21.7	28.8	55.4	57.6	30.9	85.0	82.4	87.7
Fricatives									
z	24.0	40.9	46.7	24.7	67.4	64.6	35.9	67.9	73.4
v	23.3	31.5	43.4	43.6	63.4	39.8	40.3	49.2	59.8
f	63.4	50.9	59.1	79.6	72.5	70.2	88.8	84.1	90.8
th	1.0	36.4	37.1	0.0	55.3	34.2	6.3	41.4	42.9
s	84.3	51.4	56.9	92.3	70.1	67.3	98.4	78.7	84.2
sh	64.8	37.7	49.8	59.0	58.4	62.3	76.1	74.3	82.2
dh	4.7	23.8	35.3	26.9	49.2	14.8	31.1	52.4	68.5
ch	2.2	27.3	33.1	6.5	57.5	35.0	23.6	61.8	68.6
jh	5.5	20.4	41.2	2.9	47.6	38.1	26.3	69.2	74.4
Silence									
sil	90.8	55.5	73.5	90.4	86.9	94.7	95.7	94.3	95.6

50-NN | 8-Gaussian Static | 8-Gaussian Static + Delta

Table 2: Results of the three series of experiments on the TIMIT core-test detailed by phonemes