MIXTURE SPLITTING TECHNIC AND TEMPORAL CONTROL IN A HMM-BASED RECOGNITION SYSTEM

C. Montacié, M.-J. Caraty & C. Barras

LAFORIA-IBP, Université Paris 6, CNRS-URA 1095
4 place Jussieu, 75252 Paris Cedex 5, FRANCE

ABSTRACT

In this paper, we study various technics to improve the performance, to reduce the computation cost and the required memory of a recognition system based on HMM. For the efficiency of the system, we first study the optimization of the number of HMM parameters according to training data. We experiment a temporal control of the phonetic transitions on lexical decoding task with a significant 5% improvement. Finally, a preliminary method selecting dynamically a sub-lexicon is studied in order to reduce the lexical decoding cost.

1. INTRODUCTION

Hidden Markov Models (HMM) produce the most outstanding results in the field of speech recognition. However, this technic has several limitations such as large amount of computation cost and memory request, and the difficulty to represent the duration in hidden Markov models.

For the efficiency of the system, we first sudy the optimization of the number of HMM parameters according to training data. It consists in an iterative procedure to choice the number of Gaussian mixtures used to represent the HMM states output distributions. An objective criterion based on the χ^2 distance is used.

To represent the phonetic duration, we propose a temporal control of the inter-models transitions. The phonetic transitions are constrained using the discontinuities detected with the Forward-Backward Divergence (FBD) method.

Finally, in order to solve the problem of the memory request by the language models such as n-grams, a preliminary method to select dynamically a sub-lexicon is presented.

2. HMM TRAINING

In continuous HMM, it is usual to represent the state output distribution by a Gaussian mixture density. The distribution of the observations (i.e., the speech analysis vectors) is represented by a weighted sum of Gaussian probability densities. The choice of the optimal number of mixtures is generally guided by heuristics.

To fix a number for each HMM state, we propose an iterative procedure taking into account an objective criterion of the well-fitted representation of the training data. The criterion is based on the χ^2 distance.

2.1. Coefficients Distribution

If a set of vectors has a multi-variate Gaussian distribution, each vector component has a Gaussian distribution. The converse is wrong. As it is shown in the figure 1, the analysis vectors are not well represented by a multi-variate Gaussian distribution because one component (e.g., ΔC1 the first differential Mel frequency cepstrum coefficient) has not a gaussian distribution.

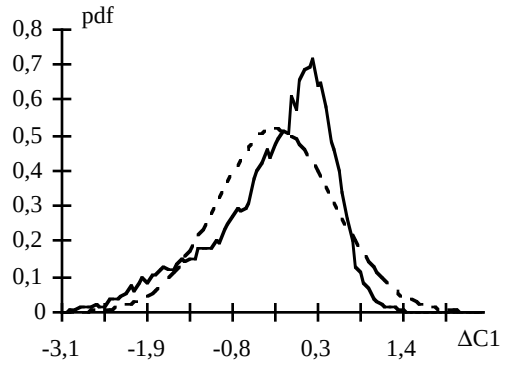

Figure 1 : Histogramm (continuous line) and estimated Gaussian probability density (dotted line) of the ΔC1 component distribution. The observations are relative to the final state of a 3-states Bakis model trained on the phonemes [iy] of the TIMIT database.

2.2. χ^2 Distance

The χ^2 distance [8] measures the dissimilarity between a probability density function (pdf) P and a set of vectors $\{y_i\}(i=1,.., N)$. At first, the representation space is divided into k clusters $\{C_i\}(i=1,..,k)$. 0,For each cluster C_i, the output probability P_i and the vector frequency f_i are computed :

$P_i = \int_{C_i} P(v) \cdot dv$, $f_i = N_i / N$, where N_i is the number of vectors belonging to the cluster C_i.

The χ^2 distance is computed by the summation on the k clusters of the quadratic substraction of the vector frequency from the probability density function:

$$\chi^2\left(\{y_i\}_{(i=1,..,N)}, P\right) = \sum_{j=1}^{k} (P_j - f_j)^2 / P_j$$

When this distance is superior to a treshold T function of the degree of freedom δ, the data probably don't follow the probability density function P. $\delta = k - 1 - N_e$, where N_e is the number of the pdf parameters estimated from the vectors $\{y_i\}(i=1,..,N)$. For instance, the degree of freedom for a Gaussian pdf and ten clusters is equal to 7 and corresponds to a treshold T of value 14. In practice, the HMM state output probability density function P is a weighted sum of multi-variate Gaussian densities. As it is expensive and difficult to divide a n-dimensional space, we decide to compute the χ^2 distance for each component of the vectors.

In the following experiment on the TIMIT database, the observations are 26-dimensional speech vectors. The components of a vector are the 12 Mel frequency cepstrum coefficients $\{C_i\}(i=1,..,12)$, the energy parameter E, and their respective differential $\{\Delta C_i\}(i=1,..,12)$ and ΔE. For each phoneme and for each state of a 3-states Bakis model, we compute the χ^2 distance between the set of vectors contributing to the considered HMM state and the output Gaussian density deduced from this training vectors (k=10, δ=7, T=14). The table 1 gives the χ^2 values computed from the training data of the phonemes [iy], for the 26 components according to the three states of the phoneme hidden Markov model. We remark only eight components are well represented by a Gaussian probability density function.

	St. 1	St. 2	St. 3		St. 1	St. 2	St. 3
C1	**5**	25	466	ΔC1	2790	83	2306
C2	28	30	41	ΔC2	159	105	56
C3	80	22	33	ΔC3	62	75	17
C4	142	397	499	ΔC4	44	144	93
C5	57	28	79	ΔC5	30	105	40
C6	**9**	19	38	ΔC6	100	74	56
C7	27	41	21	ΔC7	29	77	**1 2**
C8	31	29	36	ΔC8	15	100	28
C9	37	22	25	ΔC9	69	37	**8**
C10	22	63	31	ΔC10	**1 4**	61	**1 3**
C11	159	629	198	ΔC11	106	67	55
C12	**7**	19	15	ΔC12	56	75	**1 2**
E	38	97	280	ΔE	712	277	2100

Table 1 : For each HMM state, the χ^2 value per component. Computation from the training data of the phonemes [iy].

2.3. Mixture Splitting Procedure

An iterative procedure [2] is proposed to optimize the representation of HMM states output probabilities from training data. For an iteration, let $b(y_i)$ be the state output probability of the vector y_i computed from g Gaussian mixtures distribution :

$b(y_i) = \sum_{k=1}^{g} g_k \mathcal{G}(y_i, \mu_k, \Sigma_k)$, where g_k, μ_k , Σ_k are the HMM parameters trained by the Viterbi algorithm.

In order to get a better model, the original mixtures are splitted if they are not well representative of the data. In this purpose, we decide a first splitting criterion. A Gaussian distribution $\mathcal{G}(\mu_k , \Sigma_k)$ is not well representative of the training data if the average value $\overline{\chi^2}$ of the χ^2 distances computed from all the components is superior to the treshold T :

$$\overline{\chi^2} = \frac{1}{d}\sum_{k=1}^{d} \chi^2\left(\{y_i\}_{(i=1,..,N)}, \mathcal{G}(\mu_k, \Sigma_k)\right) > T$$

A second splitting criterion concerns the amount of data. A minimal number N_{min} of speech vectors has to contribute to the computation of the original mixture to split. Indeed, let $\tilde{\mu}$, $\tilde{\sigma}^2$ be the estimations of a Gaussian distribution $\mathcal{G}(\mu, \sigma^2)$ from n observations, the 95% confidence interval is as folowing :

$$\begin{cases} P(|\tilde{\mu} - \mu| \le 1{,}96\frac{\sigma}{\sqrt{n}}) \ge 0{,}95 \\ P(|\tilde{\sigma}^2 - \sigma^2| \le 1{,}96\frac{\sqrt{2}\sigma^2}{\sqrt{n-1}}) \ge 0{,}95 \end{cases}$$

Each Gaussian mixture $\mathcal{G}(\mu_k , \Sigma_k)$ satisfying the two previous criterion is splitted as follows :

$$g_k \mathcal{G}(\mu_k, \Sigma_k) \longrightarrow \begin{cases} \frac{g_k}{2}\mathcal{G}(\mu_k - 0{,}2\,\sigma_k, \Sigma_k) \\ \frac{g_k}{2}\mathcal{G}(\mu_k + 0{,}2\,\sigma_k, \Sigma_k) \end{cases}$$

Before the next iteration, the hidden Markov model is re-estimated. At the first iteration, a single multi-variate Gaussian distribution is trained per HMM state. The last iteration occurs when not any mixture satisfy the two splitting criteria.

2.4. Experiments

These experiments on the TIMIT database consist in HMM-based acoustic-phonetic decoding. For each of the 48 phonemes, a 3-states Bakis model is trained. N_{min}=500 is chosen. The results are the core-test phonemes recognition rates. The mixture splitting procedure previously described is generalized to the Baum-Welch algorithm taking into account the trained weights. A reference mixture splitting procedure [12] is compared. The tables 2 and 3 give for various number of Gaussian mixtures per state, the number of HMM parameters and the phonemes recognition rate.

I	Number of mixtures per state	Total number of mixtures	HMM parameters in thousands	Phonemes recognition rate (%)
1	1	144	7,5	53,2
2	2	288	15,0	56,8
3	4	576	30,0	59,2
4	8	1152	60,0	61,2

Table 2 : Reference mixture splitting procedure

I	Number of mixtures per state	Total number of mixtures	HMM parameters in thousands	Phonemes recognition rate (%)
1	1	144	7,5	53,2
2	≤2	280	14,6	57,0
3	≤4	482	25,1	58,9
4	≤8	703	36,6	61,4
5	≤16	845	43,9	61,4

Table 3 : χ^2-based mixture splitting procedure.

The figure 2 represents the phonemes recognition rates as a function of the number of HMM parameters for each mixture splitting procedure. As the results show it, the estimation of the state output probability using the χ^2-based procedure, comparatively to the reference procedure, allows a reduction of the number of HMM parameters by 50% without decreasing the phonemes recognition rates. The same procedure gives a 69,1% recognition rate when 404 contextual HMM models (190000 parameters) are trained [2].

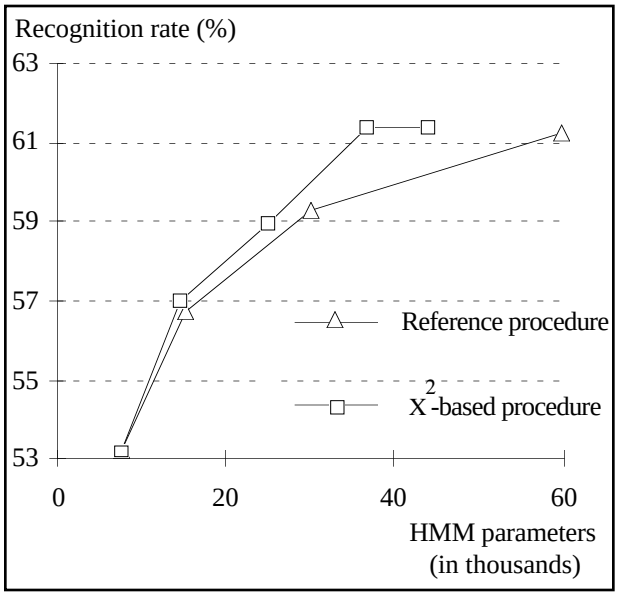

Figure 2 : Phonemes recognition rates as a function of the number of HMM parameters.

3. HMM DECODING

One major weakness of HMM is their exponentially decreasing duration law, which is unrealistic for speech. It is possible to correct the initial probability with more appropriate duration laws [7][6], but these a priori models don't make use of the acoustic signal during the decoding. The signal temporal dynamic can also be partially taken into account with the differential spectral coefficients [9], which are now widely used. But this kind of information is very implicit. We propose in this study a more precise temporal control : we constrain the phonetic transitions in the decoding system, using the discontinuities detected with the Forward-Backward Divergence (FBD) method [1]. An analysis of these observed discontinuities shows a high correlation between the temporal distribution of the discontinuities and the instants of phoneme transitions. We intend to use this segmental information in a HMM-based system. The inter-model transition probability is modified according to the detected discontinuities.

3.1. HMM with Constrained Transitions

A recognition system based on continuous HMM is applied to continuous speech recognition. The discontinuities detected with the FBD method are used in the decoding system for the control of transitions between the phonetic models. In the identification step using the Viterbi algorithm, the transition probability $P(M_j/M_i)$ between the phone models M_i and M_j depends on the distance of the nearest discontinuity. A distance higher than an a priori threshold D_m will induce a decrease of the inter-models transition probability with a factor α.

3.2. Lexical Decoding

In a previous study [3], the temporal control of the phonetic transitions have been tested for a phonetic decoding task without a significant improvement. However we test this technic for a more complex task : the lexical decoding. For these experiments, the TIMIT database including 6107 words is used. 404 contextual HMM models are trained. Each word is described by the concatenation of contextual phonetic models. The language model is made from the words pairs trained from the 2342 TIMIT sentences. There are 192 test-sentences with 1570 words. Several Log(α) factors between 0 and -8 are tested. The treshold D_m is equal to 2 cs. The results are given in the table 4. HMM unconstrained transitions correspond to Log(α)=0. We remark that the temporal control reduces the word insertion rate and increases the word identification rate.

Log(α)	0	-2	-4	-6	-8
Accuracy	85,7%	86,3%	87,6%	87,3%	87,2%
Substitution	13,4%	12,9%	11,7%	12,0%	12,2%
deletion	0,9%	0,8%	0,7%	0,7%	0,6%
Insertion	12,1%	10,9%	9,0%	9,0%	8,9%
Recognition	73,5%	75,5%	**78,6%**	78,2%	78,3%

Table 4 : Lexical decoding rate on the TIMIT core-test. Constrained transitions influence.

4. DYNAMIC LEXICON

Many HMM-based recognition systems use a one pass algorithm for the lexical decoding. Several resources (e.g., acoustic, lexical, language) have to be simultaneously present in the computer memory. With large vocabularies, it becomes important to determine a short list of candidate words [5] [10] (i.e., a sub-lexicon) before computing slow and detailed acoustic, lexical and language matches [4]. To use such a sub-lexicon, we propose a 3-steps test-sentence decoding (phonetic decoding, word hypotheser, lexical decoding) described in the figure 3. The computation cost is inchanged, because the output probabilities are stored and the memory request is lower. The performance of this approach depends on the cover of the test-sentence by the sub-lexicon.

4.1. Dynamic Lexicon Selection

A sub-lexicon is selected for each test-sentence. The selection procedure is based on a dissimilarity measure between a phonetic transcription of a word and the acoustic-phonetic decoding of the test-sentence. This measure is computed by the Wagner and Fisher algorithm [11]. The confusion cost matrix is obtained from the acoustic-phonetic decoding of the training-sentences.

Preliminary experiments show the only use of the normative phonetic transcription doesn't allow a selection of a sufficient cover of the test-sentence by a small size sub-lexicon. We choose to extract the pronunciation variants from the database labeling. This extraction uses an iterative time warping between the words of a training-sentence and its phonetic labeling.

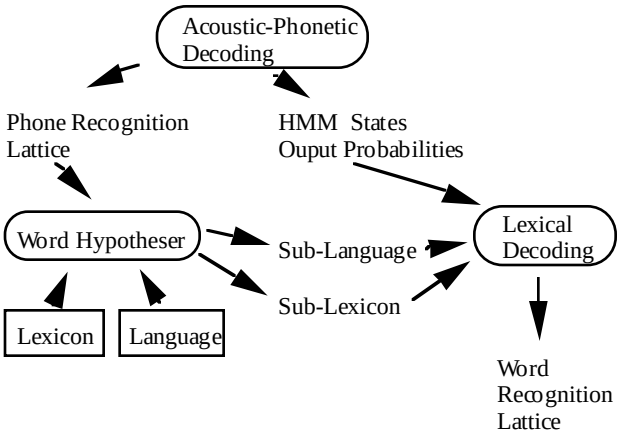

Figure 3 : A 3-steps speech decoding using dynamically a selected sub-lexicon.

4.2. Experiments

The test-corpus is composed of 1344 TIMIT test-sentences (i.e., 10980 words). Three kinds of word phonetic variants are tested : the normative phonetic transcription P_1, the phonetic variant extracted from the labeling of the training-set P_2, the phonetic variant extracted from the labeling of the training-set and the test-set P_3. The pourcentage of the test-sentences cover is computed for two sub-lexicon sizes : 10% and 20% of the original lexicon size (i.e., 600 and 1200 candidates words).

In the table 5, we remark the normative phonetic transcription is insufficient to select a dynamic lexicon based on a test decoding. The word phonetic variants are an essential resource for a dynamic lexical-based decoding.

Sub-lexicon size	600 (10%)	1200 (20%)
P_1	79,2%	87,0%
P_2	87,0%	92,5%
P_3	91,0%	95,4%

Table 5 : Test-words cover as a function of the sub-lexicon size.

5. CONCLUSIONS

This paper show the χ^2-based mixture splitting technic improves the ratio between required memory and performances for the hidden Markov models training. For the same number of HMM parameters, a 2% improvement of phonetic decoding rate is obtained. For the same performance, the number of HMM parameters can be reduced by 50%. The lexical decoding can be efficiently controled by segmental information such as the discontinuities detected by the FBD method. This temporal control improves of 5% the lexical recognition rate. A 3-steps algorithm using dynamic lexicon will be used to process a large vocabulary recognition.

6. REFERENCES

1. R. André-Obrecht. *A New Statistical Approach for the Automatic Segmentation of Continuous Speech Signals*. IEEE Trans. IEEE Trans. ASSP, vol. 36, no. 1, pp. 29-40, 1988.

2. C. Barras. *Reconnaissance de la parole continue : adaptation au locuteur et contrôle temporel dans les modèles de Markov cachés.* Phd Thesis, 1996.

3. C. Barras. *Temporal control and training selection for HMM-based system.* Eurospeech, pp. 27-30, 1995.

4. M.-J. Caraty, C. Barras, F. Lefèvre & C. Montacié. *D-DAL : un système de dictée vocale developpé sous l'environnement HTK*. 21èmes JEP, 1996.

5. L. Fissore, P. Laface, G. Micca & R. Pieraccini. *Lexical Access to Large Vocabularies for Speech Recognition.* IEEE TASSP, vol. 37, n° 8, pp. 1197-1213, 1989.

6. S.E. Levinson. *Continuously Variable Duration Hidden Markov Models for Automatic Speech Recognition*. Computer, Speech & Language, vol. 1, no. 1, pp. 29-45, 1986.

7. M. J. Russel & R. K. Moore. *Explicit Modelling of State Occupency in Hidden Markov Models for Automatic Speech Recognition*. Proc. ICASSP, pp. 5-8, 1985.

8. G. Saporta. *Probabilités : analyse des données statistiques.* Technip, 1990.

9. F. K. Soong & A. E. Rosenberg. *On the Use of Instantaneous and Transitional Spectral Information in Speaker Recognition*. Proc. ICASSP, pp. 877-880, 1986.

10. C. Waast, L. Bahl & M. El-Bèze. *Fast Match on Decision Tree*. Eurospeech, pp. 909-912, 1995.

11. R. A. Wagner & M. J. Fisher. *The String to String Correction Problem*. JACM, 1974.

12. S.J. Young. *HTK Version 1.4: Reference Manual and User Manual*. Cambridge University Engineering Department - Speech Group, 1992.