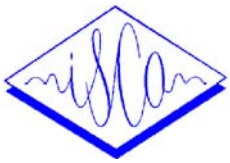


4th International Conference on Spoken Language Processing (ICSLP 96)
Philadelphia, PA, USA
October 3-6, 1996

DURATION MODELING WITH EXPANDED HMM APPLIED TO SPEECH RECOGNITION

Antonio Bonafonte, Josep Vidal and Albino Nogueiras
{antonio,pepe,albino}@gps.tsc.upc.es

Universitat Politècnica de Catalunya
c/Gran Capità s/n
08034 Barcelona (SPAIN)

ABSTRACT

In this paper, the occupancy of the HMM states is modeled by means of a Markov chain. A linear estimator is introduced to compute the probabilities of the Markov chain. The distribution functions (DF) represents accurately the observed data. Representing the DF as a Markov chain allows the use of standard HMM recognizers. The increase of complexity is negligible in training and strongly limited during recognition. Experiments performed on acoustic-phonetic decoding shows how the phone recognition rate increases from 60.6 to 61.1. Furthermore, on a task of database inquires, where phones are used as subword units, the correct word rate increases from 88.2 to 88.4.

1. INTRODUCTION

Hidden Markov Modeling (HMM) techniques have been applied successfully to speech recognition problems. However, it has been claimed [1-7] that a major weakness of HMM is that the probability density functions (DF) which model the duration of the states are assumed to be exponential, which are not appropriate for modeling the speech events which are characterized by HMM states. In order to cope with this deficiency some authors have proposed to model explicitly the state duration. In these models the first order Markov hypothesis is broken in the loop transitions. Thus, the new models have been called Hidden Semi-Markov Models (HSMM).

The first idea, up to the authors knowledge, is due to Fergurson [1] and consists in explicitly define a DF per state, p_i, which controls the occupancy in each state. In his paper, Fergurson estimated $p_i(d)$ from training data. As the number of parameters increases by D (being D the maximum duration in any state), an enormous database is required to accurately estimate the models. Fergurson himself suggested the possibility of using parametric DF for reducing the number of parameters.

Levinson [4] extended the Baum-Welch algorithm and proved its convergence for parametric HSMM. He also gave the details for the case of choosing the Gamma function. Other functions have been proposed in [2,5,6].

This research was supported by CICYT under contract TIC95-0884-C04-02

Apart from the number of parameters to be estimated, one problem with HSMM, is the increase on the computational complexity. In first implementations the computation increased by a factor D^2. In fact, more efficient algorithms can be used so that the computation increases by a factor D. The computation can still be reduced using some propierties which are accomplished by the used parametric DF [7]. Results reported on [7] show how the computation increases only by a factor 3.

All the recognizers of the reviewed papers about HSMM, are extensions of the Viterbi algorithm. These algorithms compute the probability at time t based, not only on the values at t-1, but also on the values at preceding times. Therefore, the needed storage increases by a factor D. This is not important in isolated speech recognition where the size of the space search is small. However, it becomes crucial when working on continuous speech recognition with large vocabularies and complex language models. Furthermore, in these tasks, beam search is usually used and it is difficult to extend a data-driven beam search to the case of using HSMM. Some track would be needed to follow each state in the beam, at different times.

The objective of this paper is to model the duration of the states of HMM. However, as we want to use it for continuous speech recognition, the proposal has to be inexpensive in time, memory, and implementation complexity. The reviewed algorithms do not accomplish all these constrains.

Another way to treat the problem is to model implicitly the state duration by adding more states to a conventional HMM. The Fergurson model can be represented by conventional HMM where each state is substituted by a Markov chain of D states with tied observation probabilities. Rusell and Cook [3] proposed to substitute the Markov chains proposed by Fergurson by smaller and more versatile versions (also with tied observation probabilities). The idea is to use a Markov chain in order to approximate any probability function. In [3], this idea was used only to define the HMM topology before training the HMM. In this paper the parameters of the Markov chain are estimated directly from the duration data. In this way, the number of states of the Markov chain depends on the DF to be modeled.

2. ESTIMATION OF THE PDFs

Most of the methods which have been presented to model the duration choose the DF so that the speech events are properly modeled. However, in [7], different DF were tried producing

similar improvement over the results obtained without modeling the duration. In this paper, the DF is chosen so that the complexity of the recognizer does not increases very much.

2.1. DFs modeled by a Markov chain

Suppose that a random variable (RV) $\{x_t\}$ (which represents the number of frames which are spent at HMM state) is modeled by a Markov chain. For illustrative purposes we will proceed with a 3 states chain (fig. 1) and afterwards we will generalize to the most general case of N states. The successive probabilities of remaining in the Markov chain are:

$$\begin{aligned}
&f_3(0) = 0\\
&f_3(1) = \beta_1\\
&f_3(2) = \alpha_1\beta_1 + (1-\alpha_1-\beta_1)\beta_2\\
&f_3(3) = \alpha_1^2\beta_1 + (\alpha_1+\alpha_2)(1-\alpha_1-\beta_1)\beta_2 +\\
&\quad +(1-\alpha_1-\beta_1)(1-\alpha_2-\beta_2)(1-\alpha_3)\\
&f_3(4) = \alpha_1^3\beta_1 + (\alpha_1^2+\alpha_1\alpha_2+\alpha_2^2)(1-\alpha_1-\beta_1)\beta_2 +\\
&\quad +(\alpha_1+\alpha_2+\alpha_3)(1-\alpha_1-\beta_1)(1-\alpha_2-\beta_2)(1-\alpha_3)\\
&\qquad\vdots\quad\vdots\quad\vdots\quad\vdots\\
&f_3(n) = \mathrm{X}_1^{n-1}\beta_1 + \mathrm{X}_{1,2}^{n-2}(1-\alpha_1-\beta_1)\beta_2 +\\
&\quad +\mathrm{X}_{1,2,3}^{n-3}(1-\alpha_1-\beta_1)(1-\alpha_2-\beta_2)(1-\alpha_3)
\end{aligned} \tag{1}$$

where the terms $\mathrm{X}_{1,\ldots,M}^{k}$ represent a multiple convolution of exponential responses with time constants equal to α_j, that is:

$$\mathrm{X}_{1,\ldots,M}^{p} = \sum_{i=0}^{p}\sum_{j=0}^{p-i}\cdots\sum_{l=0}^{p-i-j-\ldots}\alpha_1^i\alpha_2^j\ldots\alpha_M^{p-i-j-\ldots-l} \tag{2}$$

Thus, in general, for an N states sub-chain the DF of the duration RV is given by:

$$f_N(n) = \sum_{i=1}^{N}\beta_i\mathrm{X}_{1,\ldots,i}^{n-i}\prod_{j=0}^{i-1}(1-\alpha_j-\beta_j) \qquad \beta_N = 1-\alpha_N \qquad \alpha_0=\beta_0=0 \tag{3}$$

By Z-transforming this expression, and proceeding by induction, we obtain the characteristic function (CF), which turns out to be ARMA(N,N), with a zero at z=0. This zero accounts for the right displacement of $f_N(n)$ ($f_N(0)$ = 0).

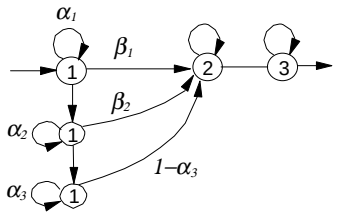


Figure 1: Expansion of a HMM of three states using a chain model. For simplicity only the first state has been expanded.

$$\Phi(z) = \sum_n f(n)z^n = z\frac{A(0)\sum_{i=0}^{N-1}b_i z^i}{\prod_{i=1}^{N}(1-\alpha_i z)} \equiv A(0)\,z\frac{\sum_{i=0}^{N-1}b_i z^i}{1+\sum_{i=1}^{N}a_i z^i} \tag{4}$$

$$\begin{aligned}
&\textit{where}\quad A(0)b_i = \sum_{j=1}^{i}(-1)^{i-j}\beta_j\Psi_{j+1,\ldots,N}^{i-j}\prod_{k=0}^{j-1}(1-\alpha_k-\beta_k)\\
&\alpha_0=\beta_0=0\\
&A(0) = 1+a_1+\ldots+a_N = \prod_{i=1}^{N}(1-\alpha_i)\\
&\Psi_{s,\ldots,N}^{0} = 1
\end{aligned} \tag{5}$$

and the term $\Psi_{1,\ldots,N}^{p}$ contains the sum of all possible products of p terms among the coefficients α_1, α_2,..., α_N. For instance,

$$\Psi_{1,2,3}^{2} = \alpha_1\alpha_2 + \alpha_1\alpha_3 + \alpha_2\alpha_3$$

The poles of the denominator must be real in order to get a positive DF. The unit integral property of the DF is satisfied with the condition of unity on the sum of the b_i.

In section 2.2 a linear method is developed to estimate the values of the terms a and b from the available duration data. Note that as far as a and b are known, then the recovery of the transition probabilities α_j is possible by rooting the AR polynomial. Afterwards, the recovery of β_j using (5) is straightforward, since β_1 depends on b_1 and, further on, β_k depends on b_k and β_1,..., β_{k-1}.

2.2. Estimation of the CF coefficients

According to (4) the CF can be expressed in the ω domain as:

$$A(\omega)\Phi(\omega) = A(0)\,B(\omega) \tag{6}$$

The coefficients of its Taylor series are the moments m_k of $\{x_t\}$. By continuously deriving $\Phi(\omega)$ with respect to ω, the moment are found as functions of the parameters b and a.

$$m_k = \sum_{i=1}^{N-1}i^k b_i - \sum_{j=0}^{k-1}\sum_{s=1}^{N}\binom{k}{j}s^{k-j}v_s\,m_j \qquad b_0 = 1-\sum_{i=1}^{N-1}b_i \tag{7}$$

where:

$$v_k = \frac{a_k}{A(0)} \qquad A(0) = 1+a_1+\ldots+a_N$$

Note that each moment can be expressed as a sum of terms depending only on the MA part of the CF and terms depending on the AR part. In particular, for the N=2 case, the set of equations is as follows:

$$\begin{pmatrix} m_1 \\ m_2 \\ m_3 \end{pmatrix} = \begin{pmatrix} 1 & -1 & -2 \\ 1 & -1-2m_1 & -4-4m_1 \\ 1 & -1-3m_1-3m_2 & -8-12m_1-6m_2 \end{pmatrix} \begin{pmatrix} b_1 \\ v_1 \\ v_2 \end{pmatrix}$$

A proof for the unicity of the solution and a discussion on the variance of this estimate can be found in [8]. It is worth mentioning that in general, not every RV can be model in this way. Regarding *A(z)*, only real valued poles can ensure a positive-definite DF. Another failure can be found if the computed β are not between 0 and 1. In those cases, one has to resort to ML estimation of the parameters. A non-linear estimation algorithm allows easy introduction of constraints. Assuming independent observations of the RV, the ML approach is obtained by maximizing the probability of the *L* observations with respect to the parameters as:

$$f(\boldsymbol{x};\boldsymbol{\alpha},\boldsymbol{b}) = \prod_{n=1}^{L} \left[\left[\sum_{i=0}^{M} b_i \delta(x-i) \right] \otimes (1-\alpha_1)\alpha_1^{-x} \otimes \cdots \otimes (1-\alpha_N)\alpha_N^{-x} \right]_{x=x_n} \quad (8)$$

with respect to the coefficients *b* and the poles α. The symbol $\otimes$ denotes convolution. In order to get a global minimum, the optimization process needs a first approach to the solution that can be obtained by the linear approach. ML optimization is time consuming and in practice, order reduction of the model leads to good results.

The analysis of the estimated probabilities easily indicates the number of states of the Markov chain which have to be used to accurately model each DF. Note that if $\alpha_j+\beta_j \approx 1$ then, the chain state j+1 can be supressed.

3. THE RECOGNITION SYSTEM

In last section we have shown how Markov chains can represent a family of parametric DF. The goodness of this family to model speech will be evaluated on next section. Now, the general scheme of the recognition system will be presented. As it will be shown, the method can be used to easily extend all the recognition systems which are based on HMM.

The training scheme has four stages:

1. Training of a initial set of HMM using any available algorithm (Baum-Welch, Segmental k-means, etc)

2. Mapping the training utterances into the states of the HMM. The training utterances are time-aligned against a sequence of HMM according to the phonetic transcription of the sentence. The Viterbi algorithm can be used to produce such alignment. At this stage, and for each state, data is collected showing the number of frames spent on the state each time it is visited. It can be argued that, as classic HMM are used, the mapping assumes that the distribution of the speech events is exponential. However, the same results are obtained if the transition probabilities of the models are ignored and the alignment is based on acoustic similarities of the speech frames, without assuming any duration model.

3. For each state, a parametric DF is estimated according to section 2. The DF is represented as a Markov chain. The number of states of each chain depends of the distribution of the time spent at each state of the initial HMM set.

4. Each state of the initial HMM is expanded by the Markov chain which has been estimated for that state. All the states of the chain share the same probability distribution to model the speech features. The final HMM set has more states than the initial HMM but with the same number of observation probability functions.

The computational effort of the training algorithm is due to the estimation of the initial HMM. In some training algorithms the stage 2 is a intermediate result of stage 1. Stages 3 and 4 require negligible computation in comparison with stage 1.

Because the result of the training algorithm is a HMM set, no modification has to be introduced on the recognition algorithm. The only thing is that, in order to speed up the computation, the probabilities of tied states has to be evaluated only once. Furthermore, before making transitions between models, the best ending state has to be chosen.

As it is shown in next section, the number of states needed to model duration is around twice the number of states of the original models. As the number of observation probabilities is the same, the computational burden of this algorithm is only slightly higher than if the initial HMM set is used.

4. EVALUATION

The initial HMM set consists of 25 phone models plus a silence model. These models have been trained from 842 phonetically balanced sentences, as referenced in [9]. The acoustic features consist of 12 mel-cepstrum coefficients, the first and second derivative and the first power derivative. The models have four states, with skip transition of one state (Bakis model).

To collect data of the number of frames which are spent at each HMM state, the training sentences are aligned against the HMM models. From these data, a Markov chain is obtained for each state. The number of states of the Markov chains is distributed as follows:

- 48 states are correctly modeled with a single state
- 24 states need a Markov chain of two states to model the duration
- 32 states need a Markov chain of three states

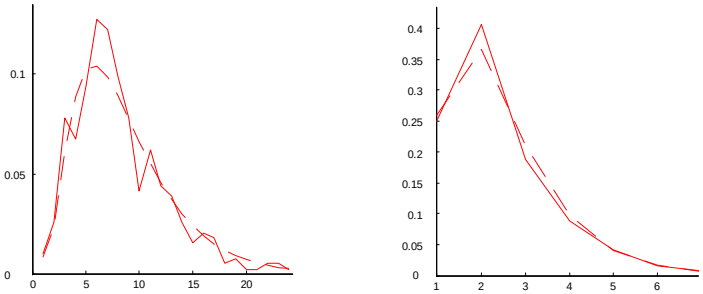
Therefore, the number of states of the final HMM set is 192, less than twice the number of states of the original set.

Figure 2.*a*, shows the histogram and the estimated DF of the occupancy of the second state of the HMM which represents sound */a/*. The DF is represented by a Markov chain of three states. Note how, with only three states, the DF fits the data very conveniently. Furthermore, it can be observed how the distribution is far from being exponential. Figure 2.*b* shows the estimation for the first state of the */tS/* model. In this case, only two states are required to represent the DF.

This models have been tried in speech recognition in two different tasks. First, in acoustic phonetic decoding: 225 phonetically balanced sentences have been recognized as a sequence of phones, without lexical information. In the second task, 600 geographical inquires [9] are recognized using a *trigram.* In both cases, the test is speaker independent and vocabulary independent.

The results show how modeling of duration improves only slightly the recognition performance. In the first experiment the percentage of correct phones (with respect to the number total of phones plus insertions) increases from 60.6 to 61.1. In the second test, the percentage of correct words increases from 88.2 to 88.4.

Some experiments have been performed to analyze the cause of this poor improvement. For instance, the duration information has been weighted in front of the acoustic information in order to increases the influence of the duration information. However, only slight improvement was achieved. In another experiment, the training corpus has been recognized in order to detect unmatching conditions between the training and test sets, but the results are similar. Finally, the transition probabilities of the final HMM have been reestimated using the Baum-Welch algorithm but no improvement has been obtained.



a) *b)*

Figure 2: Data histogram (solid line) and estimated DF (dotted line) of the occupancy of the second state of the */a/* HMM *(a)* and the first state of the */tS/* HMM *(b).*

5. CONCLUSIONS

In this paper we have proposed a method to model the duration of the states in HMM. Being the main interest of the authors the recognition of continuous speech, the objective when developing this new method was to obtain an accurate modeling of the states without increasing the computational complexity and the memory required. The algorithm proposed achieves this objective in both training and recognition. Furthermore, the method can be immediately incorporated to most of the recognition systems which are based on HMM.

The observation of the duration DF shows how with 2 and 3 states, the Markov chains can adjust very well the distribution of the duration of speech events. Unfortunately, this improvement on the model influences only slightly on the recognition results. In most of the references reviewed, improvements are reported for words. but not for subword units. The reason can be that duration modeling benefits only when the duration at each state takes large values. Some experiments with connected digits will be done on the near future to confirm this hypothesis.

8. REFERENCES

1. J.D.Fergurson, "Variable Duration Models for Speech", *Proc. of Symposium. on the Application of Hidden Markov Models to Text and Speech*, pp. 143-179, October, 1980

2. M.J.Russell and R.K.Moore, "Explicit modeling of state occupancy in hidden Markov models for automatic speech recognition," in *Proc. of ICASSP'85*, pp.5-8. Tampa, Mars, 1985

3. M.J.Russell and A.E.Cook, "Experimental Evaluation of duration modeling techniques for Automatic Speech Recognition," in *Proc. of ICASSP'87*, pp.2376-2379.

4. S.E.Levinson, "Continuously variable duration hidden Markov models for automatic speech recognition," *Computer, Speech and Language*, vol 1, pp. 29-45, Mars 1986

5. A.Falaschi, "Continuously Variable Transition Probability HMM for Speech Recognition," in *Speech Recognition and Understanding*, P.Laface and R. De Mori, Ed. Springer-Verlag Berlin Heidelberg, 1992, pp. 125-130.

6. Hung-yan Gu, Chiu-yu Tseng and Lin-shan Lee, "Isolated-Utterance Speech Recognition Using Hidden Markov Models with bounded State Duration," *IEEE Trans. on Signal Processing*, vol. 39, no. 8, pp. 1743-1751. August, 1991

7. A.Bonafonte, X.Ros and J.B. Mariño, "An efficient algorithm to find the best state sequence in HSMM", *Proc. of EUROSPEECH'93*, pp. 1547-1550. Berlin, 1993.

8. J. Vidal, A. Bonafonte, N. Fdez. de Losada, J. A. R. Fonollosa, J. Fonollosa, "Rational Characteristic Functions and Markov Chains", *IEEE Signal Processing/ATHOS Workshop on Higher Order Statistics*, pp. 226-230, Begur, Girona, June 1995.

9. A. Bonafonte, J.B. Mariño, A. Nogueiras, "Sethos: the UPC speech understanding system", *Proc. of the ICSLP´96*, Philadelphia, October, 1996.