Institute of Phonetic Sciences,
University of Amsterdam,
Proceedings 18 (1994), 111-130.

DURATIONALLY CONSTRAINED TRAINING OF *HMM* WITHOUT EXPLICIT STATE DURATIONAL *PDF*

Xue Wang

Abstract

Durational behaviour of the HMM is investigated in terms of an analytical probability density function of the whole phone model for arbitrary transitional topologies, given by the transition probabilities of *A*-parameters. Linear topology is used as an example. Based on such an analysis, the durational behaviour is manipulated by modifying the *A*-parameters in a procedure embedded in the standard Baum-Welch ML-estimation algorithm by introducing extra durational constraints on the whole-model durational statistics. The effect of such manipulation is then tested with both automatic speech recognition and segmentation, resulting in moderate improvements in performance.

1. Introduction

In order to investigate the importance of durational modelling in HMM (hidden Markov model)-based automatic speech recognition and segmentation, we analysed, as the first step, the durational behaviour of the whole HMM (instead of the single states of the HMM), without modification. It is important then to have an analytical relation between the parameters of the HMM and an expression of the durational measure, the latter can be given either in the form of a durational probability density function (pdf) or by some lower-order statistics. The second step is to modify the durational behaviour of the HMM by modifying its parameter values. Both of these two steps belong to the category of durational modelling *within* the HMM, which is the main topic of the current study. Other possible ways of durational modelling will be treated in other work. All the HMMs used in this study, including the one after modification, are HMM without the explicit state durational pdf, or called *standard* HMM in this sense. A totally different approach (the mostly used one in the literature) is to use HMM with explicit state durational pdf, or so-called hidden semi-Markov models (HSMM) (Levinson, 1986; Hochberg et al., 1993) for durational modelling. This increases the complexity of the system, and will not be studied here.

In previous work (Wang, 1993a; Wang, 1993b) we have presented the analytical form of durational pdf of a whole HMM for a special case, where the HMM topology is linear and all its selfloop probabilities are equal. General cases of left-to-right HMM with arbitrary selfloop values will be discussed in this study. Left-to-right transition topology includes all the topologies used for speech recognition; it includes any number of skipping transitions and parallel paths but no feed-back loop that involves

more than one state. A few methods of calculating the pdf will be given and some useful examples of topologies, especially the linear ones, will be discussed.

In the next step, the durational statistics collected from a set of speech data is used to modify the transition probabilities of the HMM. This is done in this study with a method embedded in the standard Baum-Welch Maximum-Likelihood (ML) training procedure, by casting extra durational constraints. However, before this procedure, a necessary step of choosing the lengths of the linear models based on the durational statistics is discussed. Finally the models trained with and without the durational constraints are compared in both automatic speech recognition and segmentation.

2. Forms of durational pdf of general HMM

2.1. Obtaining the durational pdf of the whole model

Since the total duration d in 2 states is a random variable being the sum of the duration d_1 and d_2 in two cascaded selfloops, each being an independent random variable, the pdf of a cascade of 2 different selfloops is obtained by convoluting the two geometrical pdf's (e.g. Papoulis, 1990), each being

$$P_i(d) = a_i^{d-1}(1-a_i), \qquad d \geq 1,$$

with a_i being the selfloop probability of state i.* It can be seen that the durational pdf is not a basic measure of an HMM, but a measure of the event spanning over a longer time than a single step. The principle of convolution can be easily extended to linear models composed of a cascade of $n > 2$ selfloops, and the whole pdf is

$$P_n(d) = \left[\overset{n}{*}\right] P_i(d), \qquad d \geq n,$$

where $*$ denotes convolution in d. A special case where all the selfloop probabilities are mutually different produces a relatively simple analytical form, which is a weighted sum of the individual geometrical terms, then multiplied by a constant term:

$$P_n(d) = \prod_{i=1}^{n}(1-a_i)\left[\sum_{i=1}^{n}\left(\prod_{\substack{j=1 \\ j\neq i}}^{n}\frac{1}{a_i - a_j}\right)a_i^{d-1}\right], \qquad d \geq n.$$

However, when some of the selfloops belong to some subsets with equal probability (e.g. $a_1 = a_2 \neq a_3 = a_4 = a_5$), the analytical form grows more complicated. In practice, the convolution can be performed in different ways to ease the calculation. (A direct convolution would require a troublesome bookkeeping procedure since each partial result of convolution should be further convoluted with all the remaining terms.) In the following we use z-transform for help. Assume the general case with K subsets each having n_k equal selfloops and the total $n = n_1 + n_2 + \cdots + n_K$. For simplicity we now omit the constant terms $(1-a_k)^{n_k}$ from the total pdf $P_n(d)$. The main part of the pdf is

$$\begin{aligned}\hat{P}_n(d) &= \underbrace{a_1^{d-1} * a_1^{d-1} * \cdots * a_1^{d-1}}_{n_1} * \underbrace{a_2^{d-1} * a_2^{d-1} * \cdots * a_2^{d-1}}_{n_2} * \cdots * \underbrace{a_K^{d-1} * a_K^{d-1} * \cdots * a_K^{d-1}}_{n_K} \\ &= \left[\overset{K}{*}\right]\left[\overset{n_k}{*}\right] a_k^{d-1}.\end{aligned} \tag{1}$$

* It is sufficient to use a_i instead of a_{ii} for linear models in the discussion in this section.

Its z-transform is simply

$$\hat{P}_n(z) = \prod_{k=1}^{K} \frac{1}{(z-a_k)^{n_k}}. \tag{2}$$

It is known from Wang (1993b) that each subset of n_k selfloops has a pdf of a negative-binomial form. Using z-transform properties (linearity and shift, see e.g. Rabiner et al., 1978) and some induction, we have the z-transform of the general negative-binomial terms:

$$\binom{d-1}{i-1} a_k^{d-i} \nu(d-i) \Leftrightarrow \frac{1}{(z-a_k)^i}, \qquad i = 1, 2, \cdots, n_k. \tag{3}$$

Here ν is a step function, $\Leftrightarrow$ denotes z-transformation and the binomial coefficient is

$$\binom{d-1}{i-1} = \frac{1}{(i-1)!}(d-1)(d-2)\cdots(d-i+1),$$

where $(i-1)$ is the *order* of the binomial. The z-transform of a general pdf (1) is then

$$\hat{P}_n(d) = \sum_{k=1}^{K} \sum_{i=1}^{n_k} C_k^i \binom{d-1}{i-1} a_k^{d-i} \nu(d-i) \Leftrightarrow \sum_{k=1}^{K} \sum_{i=1}^{n_k} \frac{C_k^i}{(z-a_k)^i}. \tag{4}$$

When we equal the right-hand side of (2) to that of (4),

$$\prod_{k=1}^{K} \frac{1}{(z-a_k)^{n_k}} = \sum_{k=1}^{K} \sum_{i=1}^{n_k} \frac{C_k^i}{(z-a_k)^i},$$

the problem is converted to finding the n coefficients C_k^i in the summed form by equalling the terms with the same orders of z to those in the product form. When C_k^i for each summing term is found, each term can be separately inversely z-transformed using (3) and the whole pdf is found.

To find the C_k^i we can use a classical procedure *partial fraction decomposition*. In the following, we give an example analytical form pdf whose coefficients are obtained with a *Mathematica** function *Apart*. The z-transform of the main part of this pdf is

$$\hat{P}_6(z) = \frac{1}{(z-a_1)^3(z-a_2)^2(z-a_3)},$$

namely there are 6 selfloops, in 3 subsets with 3, 2, and 1 selfloops with $a_1 \neq a_2 \neq a_3$, respectively. The pdf with the coefficients given in terms of the selfloop probabilities is**

$$\begin{aligned}\hat{P}_6(d) = {} & \frac{1}{(a_1-a_2)^2(a_1-a_3)}\binom{d-1}{2}a_1^{d-3} + \frac{-3a_1+a_2+2a_3}{(a_1-a_2)^3(a_1-a_3)^2}\binom{d-1}{1}a_1^{d-2} \\ & + \frac{6a_1^2-4a_1a_2+a_2^2-8a_1a_3+2a_2a_3+3a_3^2}{(a_1-a_2)^4(a_1-a_3)^3}a_1^{d-1} + \frac{1}{(a_2-a_1)^3(a_2-a_3)}\binom{d-1}{1}a_2^{d-2} \\ & + \frac{a_1-4a_2+3a_3}{(a_2-a_1)^4(a_2-a_3)^2}a_2^{d-1} + \frac{1}{(a_3-a_1)^3(a_3-a_2)^2}a_3^{d-1}, \qquad d \geq 6.\end{aligned}$$

---

* *Mathematica* is a software package of Wolfram Research, Inc. for both symbolic and numerical calculation on computers.

** In this formula, the legal scopes (in which the individual terms are non-zero) all begin at $d_0 < 6$, as obtained from inverse z-transforms for the terms. However the values from all the terms compensate to zero for all the points $d < 6$. Therefore the total legal range is $d \geq 6$.

The whole pdf is then simply

$$P_6(d) = (1-a_1)^3(1-a_2)^2(1-a_3)\hat{P}_6(d).$$

It can be seen that, except for the highest order binomial term in each subset, the coefficients for even such a simple cascade is complicated (the coefficient for the lowest order binomial contains 6 terms in its numerator). For another example with a cascade of a total of 10 selfloops in 4 subsets with 4, 3, 2 and 1 equal selfloops, respectively, the most complicated coefficient contains 54 terms in its numerator, as found with *Mathematica* in one hour of time for symbolic manipulation on a Macintosh machine. Reading such an analytical form pdf would not be very insightful nor pleasant. Therefore, it would not be very useful to go on with the analytical form of pdf. We end up here with a knowledge about the total number of, and the order of these negative-binomial, terms. That is, for a linear cascade, each subset of n_k equal selfloops (no matter where these selfloops are located in the cascade) with probability a_k will generally give rise to n_k negative-binomial terms with orders $i = 0, 1, \cdots, n_k - 1$, respectively (order 0 is actually a geometrical term). The whole pdf of the cascade is obtained by multiplying to the weighted sum of these terms a product

$$\prod_{k=1}^{K}(1-a_k)^{n_k}, \tag{5}$$

of the probabilities of going out of each state in the cascade. The weighting coefficients for the negative-binomial terms have a general form

$$C_k^i = \frac{N_k^i}{\prod_{\substack{l=1 \\ l\neq k}}^{K}(a_k - a_l)^{n_l+(n_k-i)}},$$

where the numerator N_k^i has an irregular form as seen in the previous example. Note that for the very extreme case where all the n selfloop probabilities in the cascade are the same, the analytical form reduces to the simplest one, being just one binomial term of order $n-1$.

In the aforementioned discussion for single cascade and in the discussion for general left-to-right topologies in the next sub-section, the analytical form pdf is only meant to give some insight. If one is interested merely in a *numerical* form of the durational pdf, however, a simpler alternative is to make use of a property of the Markov chain (Lloyd, 1980): The probability of going from state 1 to state n of a Markov chain in exactly d time steps is an entry in a product matrix of the transition matrix A, namely

$$P_n(d) = \hat{a}_{1n}, \qquad \{\hat{a}\} = \hat{A} = A^d,$$

where A^d denotes the multiplication of A to itself for d times. This makes one point on the pdf. The whole pdf can be calculated for all different d values under concern.

2.2. Analysis of whole-model pdf

The full-pdf of any left-to-right HMM can be obtained by considering each linear path (each distinct route going from the beginning to the end of the whole model) separately as above, and summing them together with weights as in (5) for that path. Each path has a *legal scope* within which the pdf is non-zero and is usually $d \geq d_0$. In a more general case where some states in a path have no selfloop, the pdf for this path will not contain binomial terms for those states. The contribution of these states is simply a

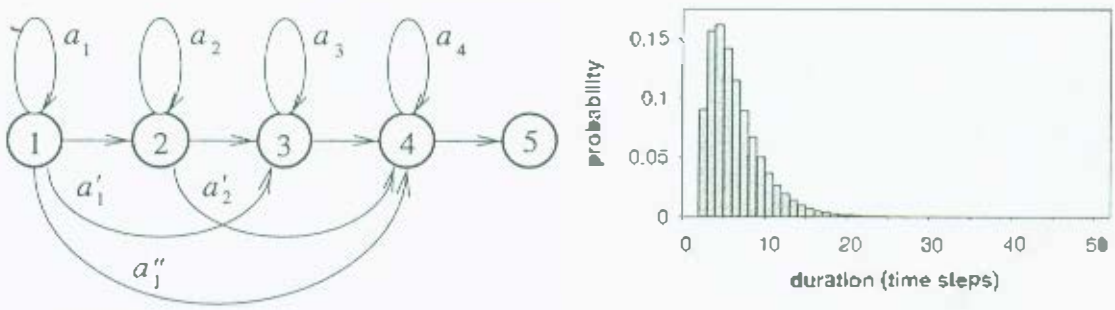

Figure 1. Left: an HMM with its transition probabilities shown. It has 5 states and 4 paths. These paths given in their state-indices are (1→2→3→4→5), (1→2→4→5), (1→3→4→5) and (1→4→5) respectively. Right: its durational pdf with specific values (see Table 1) assigned to all the transition probabilities, as a function of any constant time steps.

constant that should be included in (5). In the following, we first give a pdf of a particular HMM as an example. Its transition topology is shown in Figure 1, and the selfloop probabilities are all different.

The complete pdf of this HMM consists of 4 terms, each concerns a linear path. Although the geometrical terms with the same a_i but from contributions of different paths, can be put together, they may have different legal scopes of d. These scopes are indicated on the right side of the lines for separate paths, in the following formula:

$$P_4(d) = (1-a_1-a'_1-a''_1)(1-a_2-a'_2)(1-a_3)(1-a_4)$$

$$\left[\frac{a_1^{d-1}}{(a_1-a_2)(a_1-a_3)(a_1-a_4)} + \frac{a_2^{d-1}}{(a_2-a_1)(a_2-a_3)(a_2-a_4)}\right.$$

$$\left. + \frac{a_3^{d-1}}{(a_3-a_1)(a_3-a_2)(a_3-a_4)} + \frac{a_4^{d-1}}{(a_4-a_1)(a_4-a_2)(a_4-a_3)}\right] \qquad (d \geq 4)$$

$$+a'_1(1-a_3)(1-a_4)\left[\frac{a_1^{d-1}}{(a_1-a_3)(a_1-a_4)}\right.$$

$$\left. + \frac{a_3^{d-1}}{(a_3-a_1)(a_3-a_4)} + \frac{a_4^{d-1}}{(a_4-a_1)(a_4-a_3)}\right] \qquad (d \geq 3)$$

$$+(1-a_1-a'_1-a''_1)a'_2(1-a_4)\left[\frac{a_1^{d-1}}{(a_1-a_2)(a_1-a_4)}\right.$$

$$\left. + \frac{a_2^{d-1}}{(a_2-a_1)(a_2-a_4)} + \frac{a_4^{d-1}}{(a_4-a_1)(a_4-a_2)}\right] \qquad (d \geq 3)$$

$$+a''_1(1-a_4)\left[\frac{a_1^{d-1}}{a_1-a_4} + \frac{a_4^{d-1}}{a_4-a_1}\right]. \qquad (d \geq 2)$$

This looks complicated. But if we give some specific values to each a_i, such as the ones given in Table 1,

Table 1. Specific values of transition probabilities for the example.

a_1	a'_1	a''_1	a_2	a'_2	a_3	a_4
0.1	0.2	0.3	0.4	0.5	0.6	0.7

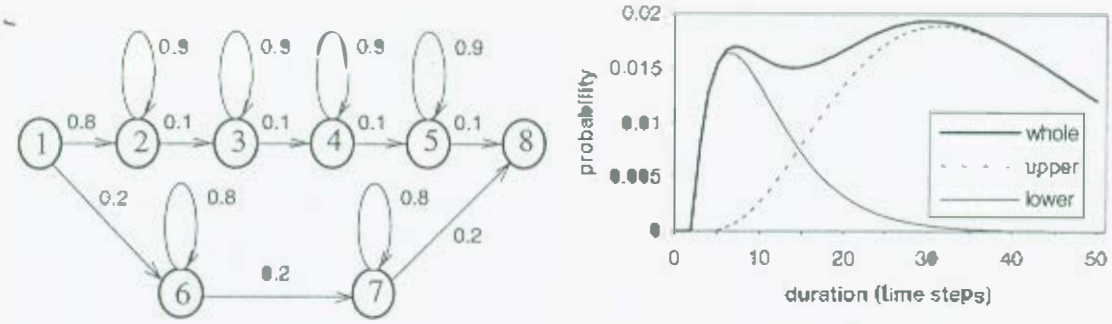

Figure 2. Left: an HMM with 2 parallel paths each containing selfloops. Right: the durational pdf of the upper and lower paths and that of the whole model.

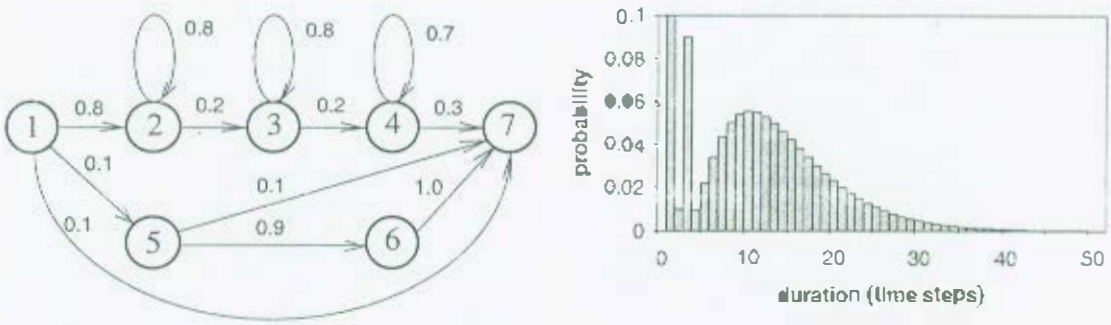

Figure 3. Left: the Kai-Fu Lee model with transition probabilities shown. The 3 lower paths given in their state-indices are (1→7), (1→5→7) and (1→5→6→7), respectively. Right: its durational pdf. The values at duration 1, 2 and 3 are contributions from the 3 lower paths respectively, all without self.oops.

then the total pdf , after being organised for different scopes of d, looks simple:

$$P_4(d) = \begin{cases} 0, & d < 2; \\ 0.09, & d = 2; \\ 0.156, & d = 3; \\ 0.21(0.1)^{d-1} - 0.4(0.4)^{d-1} - 0.96(0.6)^{d-1} + 1.15(0.7)^{d-1}, & d \geq 4, \end{cases}$$

and this is plotted in Figure 1.

Below we give two examples of models with parallel paths, both with actual values of all the transition probabilities. The first model (Figure 2)* consists of two paths, each contributing a single peak and the pdf of the whole model shows two peaks. The second model (Figure 3) is the 'Kai-Fu Lee' model (Lee, 1989), with four paths. Each of the 3 lower paths contributes only a single point on the pdf (the first 3 points) since they do not contain any selfloops.

As observed in the above examples and as so far always has been true in our practice with real data, the pdf's of linear models and the model with skipping paths have always a single peak, regardless of whether the selfloop probabilities are equal or not. (This has not yet been proven theoretically, though.) This can be regarded as a general behaviour of the durational pdf of the HMM. Some later discussion for a single linear path will be based on equal-self.oops while the conclusion will be general for

* The durational pdf's in this study are plotted in different styles for the sake of clarity. However it has to be noted that even with a *continuous* line drawing, the pdf values are only defined at *discrete* time steps (they should actually be called *pmf*: probability mass functions).

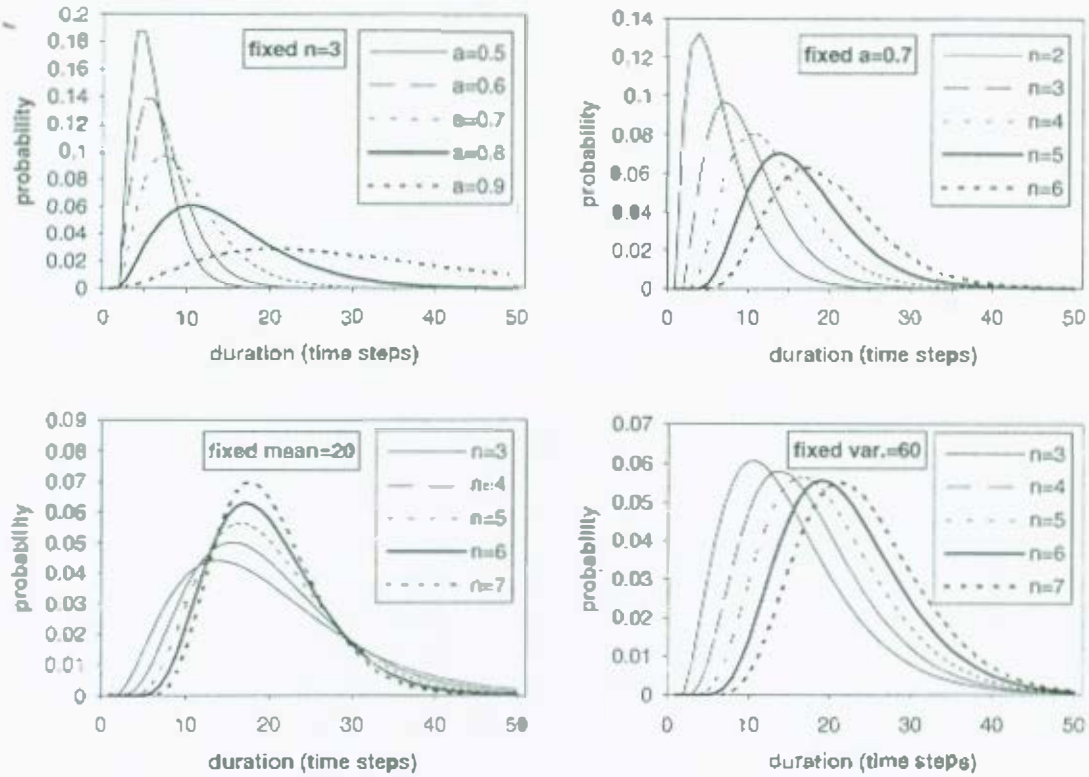

Figure 4. Durational pdf's for linear models with equal selfloops. Different panels show cases when either n, a, mean or variance is fixed while other variables are varying.

models with different selfloops. Skipping transitions provide possibility for shorter *minimal duration* than the length n of the path, which, without skipping, would be exactly n. Pdf with more than one peak can be obtained with *independent parallel* paths, namely those paths which do not share states except for the begin and the end states.

As will be seen in later sections, the real data of speech segments always show single-peak behaviour. Furthermore, using parallel paths would cause incorrectness, in that during HMM training, parameters in different paths would be trained with the *same* data while different paths should model *different* processes. Therefore in our study we do not use any parallel paths. We also found out that skipping transitions for minimal duration are unnecessary if one selects the model length n carefully according to the data (next section). Therefore we only use *linear models* with one path and no skips.

We will use some durational statistics in the data to constrain the model parameters. As a first step we choose to use durational mean μ and variance σ^2 of the speech segments. From probability theory (e.g. Papoulis, 1990) we know that, each state of the geometrical pdf has a durational μ and σ^2 as

$$\mu_i = \frac{1}{1-a_{ii}}; \qquad \sigma_i^2 = \frac{a_{ii}}{(1-a_{ii})^2},$$

and the μ and σ^2 of the whole linear model are sums of those of individual states. For the special case of equal $a_{ii} = a$, such relations between n, a, μ and σ^2 become

$$\mu = \sum_{i=1}^{n} \frac{1}{1-a_{ii}} = \frac{n}{1-a}; \qquad \sigma^2 = \sum_{i=1}^{n} \frac{a_{ii}}{(1-a_{ii})^2} = \frac{na}{(1-a)^2}. \tag{6}$$

These relations are shown in the upper panels of Figure 4*. From (6), when either n or a is fixed, both μ and σ^2 increase monotonically with the other variable. On the other hand, if we eliminate a from (6) to make it implicit and rewrite as

$$\sigma^2 = \mu - \frac{\mu - n}{n}; \qquad \mu = \frac{n}{2} + \sqrt{\sigma^2 n + \frac{n^2}{4}},$$

it can be seen that σ^2 decreases as n increases when μ is fixed, and μ increases with n when σ^2 is fixed. These dependencies are shown in the lower panels of Figure 4.

3. Constrained training of HMM embedded in ML procedure

3.1. Paradigm of the constrained training

When we include the initial probability π in the transition probability A, the whole parameter set of the standard HMM to be estimated is $\lambda = (A, B)$ where B is the observation probability. Then the auxiliary function can be decomposed into two terms for A and B respectively, and these are maximised separately. The solutions to the constrained maximisation problems give the formulae used for re-estimation of the parameters given their old values. The unity constraint used is given by the properties of the probability measures (first row in Table 2), which e.g. for A is

$$\sum_{j=1}^{n} a_{ij} = 1, \qquad i = 1, 2, \cdots, n.$$

Table 2. Different paradigms of parameter constraining in training of HMM.

HMM	constraints for A	constraints for B	constraints for dur.
standard	unity	unity	
explicit duration	unity	unity	seg. dur. statistics
constr. standard	unity, seg. dur. stat.	unity	

When the system is not standard HMM so that the parameter set contains parameters in addition to A and B, they too contribute to the auxiliary function, and should also be maximised using their own constraints. An example of such training is in Hochberg et al., (1993) using HSMM with Gaussian pdf (second row of Table 2). In our approach, since the modelled segment duration is given by the A parameters of the standard HMM, an extra constraint is used on the same parameter (third row of Table 2).

The main difference between the HMM with the explicit state duration and the HMM constrained with the durational pdf is, that the former has extra parameters with their own constraints and the latter has no extra parameters beyond the standard HMM, but the same parameters are confined with extra durational constraints.

3.2. Embedded training with extra durational constraints

Having seen the relations and dependencies between n, a, μ and σ^2, we know that for a given pair (μ, σ^2) from the data of a segment to be fitted, the choice of model length n is constrained within a range. This range is further shrunk when we use a particular

* Note that on all the pdf plots, μ is roughly related to the horizontal position of the curve, whereas σ^2 to the breadth of the curve.

way of a necessary numerical search to be discussed later. These procedures are given in both Appendix 1 and Appendix 3, and the final range for n is

$$\frac{\mu(\mu-1)+\sigma^2}{\mu-1+\sigma^2} < n < \mu+1-\sqrt{2\sigma^2+1}.$$

Once n is chosen, we can start with the Baum-Welch algorithms for Maximum Likelihood (ML) training with extra constraints. The unity constraint must still hold, only for our special case of linear models, it reduces to

$$a_{ii} + a_{i,i+1} = 1, \qquad i = 1,2,\cdots n-1,$$

and can be used to eliminate $a_{i,i+1}$. After this, the set of equations obtained using the Lagrange-multipliers concerning the A parameters are (B parameters are irrelevant)

$$\begin{cases} \dfrac{1}{\tilde{a}_{ii}} \sum_m \sum_t \gamma_{t-1}^{(m)}(i,i) - \dfrac{1}{1-\tilde{a}_{ii}} \sum_m \sum_t \gamma_{t-1}^{(m)}(i,i+1) + \\ \qquad + \theta_1 \dfrac{1}{(1-\tilde{a}_{ii})^2} + \theta_2 \dfrac{1+\tilde{a}_{ii}}{(1-\tilde{a}_{ii})^3} = 0, \quad i = 1,2,\cdots n; \\ \sum_{i=1}^{n} \dfrac{1}{1-\tilde{a}_{ii}} = \mu; \\ \sum_{i=1}^{n} \dfrac{\tilde{a}_{ii}}{(1-\tilde{a}_{ii})^2} = \sigma^2, \end{cases} \tag{7}$$

where the "counts" γ are obtained with the A and B values at the previous iteration using the usual Baum-Welch procedure, and are summed over all time t and all observation sequences m (e.g. Kamp, 1992). $\tilde{a}$ are new values to be sought after the current iteration, and θ_1 and θ_2 are two multipliers.

It turned out that this set of $(n+2)$ non-linear equations cannot be solved analytically to give formulae for calculating the new A values from old ones, as is the case in ML procedure for standard HMM without extra constraints. We have chosen to use a Newton-Raphson (Press et al., 1989) iteration procedure to find numerical solutions with some initial values (see the next section). The following set of $2n$ equations further constrain the iteration procedure to find only meaningful values of $\tilde{a}$:

$$\begin{matrix} \tilde{a}_{ii} > 0, \\ \tilde{a}_{ii} < 1, \end{matrix} \qquad i = 1,2,\cdots n.$$

The actual method of using the Newton-Raphson procedure is given in Appendix 2. In general, it searches for improvement of solutions from the points of current iteration, based on local derivatives of the set of equations including (7) and further constraints. The details of numerical search together with the necessary initial points chosen on the basis of data constraints (μ,σ^2) is given in Appendix 3.

4. Results

4.1. Results of durational pdf fitting

Before the recognition and segmentation runs, the effect of the durational constraint on the modelled durational pdf of the HMM is checked. In this study, we used the whole set of the TIMIT database (see Zue et al., 1990, and the documents included in the TIMIT CDROM) in the experiments. TIMIT contains American English continuous speech from a total of 630 speakers each reading 10 sentences (of which 2 are the

same across all the speakers). A total of 1680 sentence utterances from 168 speakers were used for testing and the rest for training (thus a speaker-independent performance). Our system is a phone-based continuous speech recogniser.

Firstly the durational histograms for all the 61 original TIMIT phones are estimated from the whole training and testing sets. The statistics pair (μ,σ^2) are calculated (and slightly modified for a few phones) and the suitable lengths n for all the phones are chosen. These n range from 3 to 10 for the whole system. Table 3 shows examples of the situations of the modelled durational statistics calculated from the HMMs trained with and without the durational constraints, for 8 phones. For easy comparison, σ instead of σ^2 is used in the table.

Table 3. Durational statistics in the TIMIT data and from the models. The first column from the left shows the symbols of 8 example phones. The second column lists the durational μ and σ directly calculated from the hand-segmented label files. The third column shows those necessarily modified data statistics. The fourth column shows the allowed range for choosing a suitable n and the actual chosen n for each phone. The last two columns show the durational μ and σ calculated from the models trained without and with the durational constraints (abbreviated as "ct"), respectively.

	original data		modified data		after data modification				model no ct		model with ct	
	μ	σ	μ	σ	$\tilde{n}_{\min}$	$n^L_{\max}$	$n^U_{\max}$	n	μ	σ	μ	σ
aw	20.45	6.44			7.20	12.27	14.48	8	19.24	5.34	20.45	6.44
b	2.19	0.89	3.50	0.83	2.96	2.96	3.03	3	3.30	0.60	3.50	0.83
ih	9.84	3.53			4.67	5.75	6.78	5	8.51	2.65	9.84	3.53
pau	23.34	15.78			2.84	2.00	8.05	3	18.64	10.65	23.34	15.71
q	8.16	3.94		3.37	3.77	4.23	5.25	4	6.25	1.92	8.16	3.37
sh	14.51	3.71			7.70	10.17	11.27	8	16.10	4.49	14.51	3.71
y	8.34	4.40		3.49	3.76	4.30	5.31	4	14.94	6.48	8.34	3.49
z	10.51	3.91			4.65	5.90	7.07	5	12.59	4.56	10.51	3.91

Note that if we use relation (6) of the previous section and force all the selfloop probabilities to be equal $a_{ii}=a$, there is no freedom in choosing n that can fit the data statistics pair (μ,σ^2): both n and a are only allowed to be a fixed value. We see this by solving n (and a) from (6) for the case of equal a:

$$n=\frac{\mu^2}{\mu+\sigma^2}, \qquad a=\frac{\sigma^2}{\mu+\sigma^2}.$$

The n here may well be a non-integer, which is then a problem. Now when we allow a_{ii} to be different, we usually get a range $(\tilde{n}_{\min}, n^L_{\max})$ in which we can choose n freely. We have chosen the smallest integer n for each phone. The determination of the values of a_{ii} will be left for the training procedure.

The following can be seen from Table 3:

1. For those phones (e.g. /b/) with $\mu<3$, the data μ was modified and the data σ was modified accordingly, in order to be able to choose an $n\geq 3$ (see Appendix 3);
2. For /q/ and /y/, since no suitable $n<n^L_{\max}$ can be found based on the original data μ and σ, the data σ were modified (decreased) (Appendix 3);
3. For both the two phones /b/ and /pau/ (between-word pause) which require an $n=3$, the worse upper limit $n^U_{\max}$ were used (true for all phones with $n=3$, see Appendix 3);
4. For those phone with $n>3$, some have a small range $(\tilde{n}_{\min}, n^L_{\max})$ in choosing n, while others have a larger one.

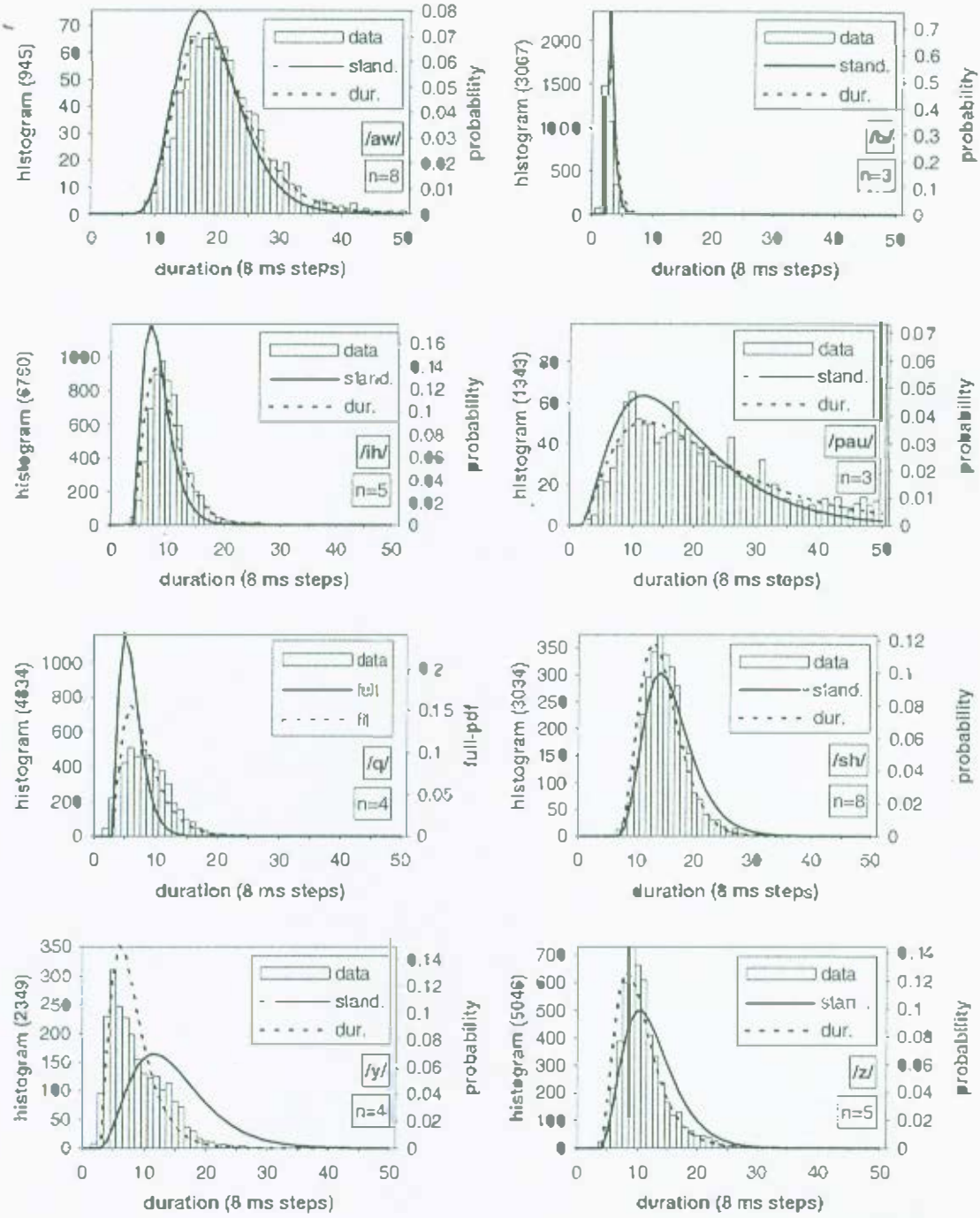

Figure 5. Examples of phone durational pdf's from the TIMIT database: Data histograms (data: blank vertical bars) are estimated from the number of instances given between the brackets. Pdf's of standard (stand.: continuous curve) and durationally constrained (dur.: dashed curve) IEMM are scaled to the histogram. Histograms are plotted against the left scale, and pdf's are plotted against the right scale, of each panel. The chosen model length n are also shown.

5. The modelled μ and σ of the models trained *without* durational constraints show various degree of deviations from the data μ and σ, the worst being /y/ (see also Figure 5);
6. The modelled μ and σ of all the models trained *with* durational constraints show good agreement with the data μ and σ.

Note that since the fitting of durational μ and σ^2 only uses these two lower-order statistics as the first approximation, the fitting of the entire pdf remain to be checked. This is shown in Figure 5. for the pdf's of these same set of 8 phones.

The first impression is that both the pdf's of the models with and without durational constraints show good fit in shape with the data histograms, indicating that the *form* (weighted sum of binomial/geometrical terms) of the pdf of linear models is suitable. This is evidence that the HSMM is unnecessary, because at the level of the whole model, a simple linear HMM can model the duration well. Furthermore, it is evident that the models trained with constraints always fit the data histograms (slightly) better as compared with the models trained without constraints. However, it can also be seen that, because of the careful choice of the model lengths, the durational pdf's without the durational constraints are already rather close to the data durational histograms (except for /y/, for which both μ and σ^2 have a very bad match with the histogram for the case without durational constraining).

The improvement in durational fitting is a clear indication of the improvement of the quality of our HMM as compared with the HMM trained with the standard Baum-Welch algorithm, at least with respect to the whole-model durational modelling accuracy. The standard Baum-Welch algorithm does not take the durational statistics as part of the training criteria, as the case in our constrained training. Therefore even with the careful choice of the model lengths, the fitting of the standard HMM in durational statistics is not *guaranteed*.

4.2. Results of recognition and automatic segmentation

The ultimate goal of durational modelling with HMM is not only to see the possibility of modelling the segmental duration accurately, but also to see if such an accurate modelling improves the performance of speech recognition and segmentation. Technically, during the durationally constrained Baum-Welch training, the constrained values of the A parameters obtained from each iteration will be used in the next iteration, in which, both the A and the B parameters will be affected by the durational constraints. (In this way B parameters are *indirectly* affected). Although it is argued that A parameters alone do not govern the transition behaviour, the new values of A and B together may have a different general behaviour (not only the transition) from that of standard HMM parameters without the durational constraints. Therefore, in recognition and segmentation, the performance may be different (hopefully improved). The effect is checked experimentally.

In our experiments of both recognition and segmentation, the whole set of TIMIT database was used. TIMIT has been used for various related research topics, e.g. speaker identification (Lamel et al., 1993a) and phone recognition (Lamel et al., 1993b).

Our recogniser uses context-independent phone models with a linear transitional topology. LPC-based cepstrum coefficients plus their time derivative and energy are used in 3 streams for the HMM. The observation probability of each state has a weighted mixture of 3 Gaussian densities, and uses a diagonal covariance matrix. The model lengths n chosen as in the previous sub-section are kept during the whole test. Then 2 setups of HMM training are performed, one with and another without the durational constraints. The same set of HMM after the initialisation training are used for both setups.

The language models used for recognition are a *regular* type of word-pair grammar and a bi-gram, both estimated from the whole testing set. Phonological rules within words are basically linear plus an optional pause at the end of each word, and

Table 4. Scores of recognition and segmentation with and without duration constraint, for speaker-independent tests on the TIMIT database. For recognition, the scores are word-correct percentages. The first score counts only substitution and deletion errors whereas the second score (between brackets) counts also insertion errors. For segmentation, the percentages are given on the correctly matched segment borders within the threshold of 20 *ms* in both directions, as compared with the hand labels.

	without duration constraint	with duration constraint
recognition	80.61% (77.73%)	86.83% (84.41%)
segmentation	83.48%	84.48%

additional closures for plosives when they do not follow immediately a silence. The "language model" for segmentation is simply an exact linear sequence of the phones in each sentence. The segmentation process is simply a "recognition" process with the identities of the phone sequence known *a priori* but the border information missing.

The scores of a preliminary experiment in recognition and segmentation are shown in Table 4. Moderate improvements are shown for both cases.

5. Conclusions

In this study, the HMM without explicit state durational pdf, but trained with constraints on the durational statistics of the acoustic segment, shows improvement in performance of both recognition and segmentation, at very little extra computation costs. The first conclusion is that the performance of the standard HMM can still be improved if extra information, such as the one about the segment duration, can be integrated into the models. Recall that the durational measures, e.g. the durational pdf or the lower-order durational statistics, are not the basic measures of the HMM. Therefore the durational information is regarded as coming from an *independent* information source, and has been brought into the models during the improved Baum-Welch training with durational constraints. Whether such integration can improve the performance of the recogniser is checked with the tests, giving us a positive answer.

The technical implementation of the integration of the durational information is achieved in three necessary steps, which are reported in depth in this paper, together with necessary mathematical development. The first step is to find the relations between the durational measures and the HMM parameters. An observation into the usual durational modelling technique using HSMM reveals (Wang, 1993b) that it is insufficient to look at the durational behaviour at the *state level* alone, and, if one looks at the whole-model behaviour, HSMM may be unnecessary, too. Then we concentrated on the durational pdf of the *whole model* of phones, and obtained the analytical forms pdf for several topologies, of the standard HMM.

In this study only the transition probabilities A are regarded as relevant. One is convinced that the standard whole HMM without the explicit state durational pdf is powerful enough to model the durational distribution of the speech segments (phones). The second necessary step is to find the suitable lengths of the HMM based on the data durational statistics. The third step is to actually constrain the A-parameter values with the data durational statistics during the Baum-Welch training. All these steps fit into a framework that the HMM both should model the process of *acoustic* observation, and should fit the whole-phone *durational* statistics well. Both goals are achieved in an integrated way using the current approach.

Some limitations of the current study are as follows. The form of the durational pdf of the HMM in this study implies that the durational behaviour to be modified is only governed by the transition probabilities A. Further research should seek for durational

pdf in which both A and the observation probabilities B play roles. The constraining technique in the current study may then be useful to *directly* modify the A and B parameters. It will be possible to get even better results than the current study in modelling and performance in recognition and segmentation. Furthermore, only the segmental duration, but no other long-term features, was used as extra information for the model improvement. Other long-term features of speech, if available from signal processing and if they can be stored in a statistical way, such as a statistical distribution of some specifications of the pitch contour, may be directly integrated using the current technique.

Appendix 1. Choice of the model length

In this Appendix we will discuss the problem of choosing the most suitable model length n based on μ and σ^2 from the actual data. The selfloop probabilities are allowed to be different in general. For convenience we take a monotonic transformation

$$u_i = \frac{1}{1-a_{ii}}, \tag{8}$$

and then the general relation in (6) is given as

$$\sum_{i=1}^{n} u_i = \mu, \qquad \sum_{i=1}^{n}\left(u_i - \frac{1}{2}\right)^2 = \sigma^2 + \frac{n}{4}. \tag{9}$$

It can be seen from this set of equations that for a given n, the set of values $\{u_i\}$ that satisfy both durational mean and variance lie on the n-dimensional circle defined by the intersection between a hyper-plane and a hyper-sphere centred at $\{½, ½, \cdots ½\}$, in the space $\Re^n$ for $\{u_i\}$.

The radius of the sphere is controlled by n and σ^2 whereas the intercepts of the plane are controlled by μ. For a given pair (μ, σ^2), different values of n correspond to 3 different situations for the intersection circle.

The first situation, corresponding to the smallest n, is given when the sphere is tangent with the plane at a single point, namely when all u (or a) are equal. By eliminating all u from (9), it is easy to verify that

$$n_{\min} = \frac{\mu^2}{\sigma^2 + \mu}. \tag{10}$$

(Note that this relation is just the same as in the last section for equal selfloop probabilities.)

The second situation applies when n (thus the sphere) is larger so that the intersection becomes a circle, but this is bounded by the condition required for $u_i > 1$ (see (8)) when a lower bound of the maximum $n^L_{\max}$ is defined*. In this situation, all the points on the circle can be used as solutions of (9). When n further increases beyond $n^L_{\max}$, some points on the circle will cause some $u_i < 1$, thus are no more solutions, and the solutions left are n disjoint pieces of arcs on the circle.

The third and extreme situation when there still are solutions is when these n arcs shrink into n isolated points. This corresponds to the largest $n = n^U_{\max}$ that can provide a solution.

We first get $n^U_{\max}$. The n points on the largest possible circle form an n-dimensional equilateral polygon. The co-ordinates of u for each endpoint of this polygon have a

* We use superscripts L and U for lower and upper bounds, respectively.

pattern that $(n-1)$ components are equal to 1 while one component is $\mu-(n-1)$. This gives e.g. $\{\mu-(n-1),1,1,\cdots,1\}$ and $\{1,\mu-(n-1),1,\cdots,1\}$. Putting any such point into the second equation of (9) we then get

$$(n-1)\left(1-\frac{1}{2}\right)^2+\left[\mu-(n-1)-\frac{1}{2}\right]^2=\sigma^2+\frac{n}{4}.$$

With a given pair (μ,σ^2), we solve for the largest possible n (for the largest sphere)

$$n_{\max}^{U}=\mu+\frac{1}{2}-\sqrt{\sigma^2+\frac{1}{4}}.$$

From now on we get $n_{\max}^{L}$ for a complete solution circle. One point on such a circle is tangent with the middle point on one of the edges of the polygon, defined by the two end points of the edge:

$$\boldsymbol{u}_x=\left\{\frac{\mu-(n-1)+1}{2},\frac{\mu-(n-1)+1}{2},1,1,\cdots,1\right\}.$$

Putting this point into (9):

$$2\left(\frac{\mu-n+2}{2}-\frac{1}{2}\right)^2+(n-2)\left(1-\frac{1}{2}\right)^2=\sigma^2+\frac{n}{4}.$$

Again, with a given pair (μ,σ^2) this gives us the 'safest' largest n with which all the points on the intersection circle are solutions for (9):

$$n_{\max}^{L}=\mu+1-\sqrt{2\sigma^2+1}.$$

From above we investigate that an ill-behaved circle given by $n_{\max}^{L}<n<n_{\max}^{U}$ * can bring a numerical searching from a solution to a non-solution point along the circle. Therefore in order to prevent numerical problems caused by this reason, we choose

$$n_{\min}<n<n_{\max}^{L} \tag{11}$$

Furthermore, it is clear that n should be an *integer,* representing the number of selfloops, and should be chosen as the smallest possible value within the region, for simplicity.

Appendix 2. Solution of non-linear equations

Generally, Newton-Raphson method searches for numerical solutions for N variables y_i given in N non-linear equations

$$f_i(y_1,y_2,\cdots,y_N)=0, \qquad i=1,2,\cdots,N. \tag{12}$$

This is achieved by using the values of the current iteration of f_i and their partial derivatives to form a set of linear equations for the local increments δy_i:

$$\sum_{j=1}^{N}\frac{\partial f_i}{\partial y_j}\delta y_j=-f_i. \qquad i=1,2,\cdots,N. \tag{13}$$

All the δy_i can be solved out using any standard method for linear equation systems, such as LU-decomposition. Then the y_i values are updated as

---

* $n_{\max}^{L}<n_{\max}^{U}$ because they both are monotonically decreasing functions of σ^2, both evaluate μ at 0 and zero-cross at $\mu(\mu/2+1)<\mu(\mu+1)$, respectively.

$$y_i^{new} = y_i^{current} + \delta y_i, \qquad i = 1,2,\cdots,N.$$

The iteration starts at some chosen initial point and ends when some convergence threshold is reached.

Our particular set of non-linear equations come from the Baum-Welch ML (maximum-likelihood) parameter estimation procedure. Firstly, the auxiliary function (e.g. Kamp, 1992) for our linear HMM with extra durational constraints in ML is

$$F = \sum_i D(i,i)\log\tilde{a}_{ii} + \sum_i D(i,i+1)\log(1-\tilde{a}_{ii}) +$$
$$+\theta_1\left(\sum_i \frac{1}{1-\tilde{a}_{ii}} - \mu\right) + \theta_2\left(\sum_i \frac{\tilde{a}_{ii}}{(1-\tilde{a}_{ii})^2} - \sigma^2\right),$$

where $\tilde{a}_{ii}$ are the new values of selfloop probabilities after the current iteration, and

$$D(i,j) = \sum_m \sum_t \gamma_{t-1}^{(m)}(i,j)$$

are the "counts" γ obtained from the previous parameter values, summed over time t and observation sequences m. Further constraints for the numerical search to be confined within the meaningful region may be written as $2n$ negative functions

$$g_k = \begin{cases} -\tilde{a}_{ii} < 0, & (k = 1,\cdots,n); \\ \tilde{a}_{ii} - 1 < 0, & (k = n+1,\cdots,2n). \end{cases}$$

Introducing some positive relaxation functions

$$s_k = x_k^2 + \varepsilon, \qquad (k = 1,\cdots,2n),$$

where $\varepsilon > 0$ is a small number to keep the computer from the edge, to bring the constraints in equation form:

$$\varphi_k = s_k + g_k = \begin{cases} x_k^2 + \varepsilon - \tilde{a}_{ii}, & (k = 1,\cdots,n); \\ x_k^2 + \varepsilon + \tilde{a}_{ii} - 1, & (k = n+1,\cdots,2n). \end{cases}$$

Now the new auxiliary function including all the constraints becomes

$$\Phi = F + \sum_{k=1}^{2n} \lambda_k \varphi_k.$$

To get the critical point for the ML of Φ we take the partial derivatives w.r.t. the $N = 5n+2$ variables, namely n of $\tilde{a}_{ii}$, 2 of θ, $2n$ of λ_k and $2n$ of x_k, respectively, and let them be zero, resulting in a total of $5n+2$ non-linear equations $f = 0$. To solve these equations we use the *linear* equations (13) about the increments δy. For clarity, we write (13) in matrix form

$$C\delta Y = -f.$$

Here $\delta Y = (\delta y_1, \delta y_2, \cdots, \delta y_{5n+2})^\tau$ (τ denotes transpose), f is the vector of $5n+2$ non-linear functions in the $5n+2$ variables, specifically,

$$f_i = \frac{1}{\tilde{a}_{ii}} D(i,i) - \frac{1}{1-\tilde{a}_{ii}} D(i,i+1) + \theta_1 \frac{1}{(1-\tilde{a}_{ii})^2} +$$
$$+\theta_2 \frac{1+\tilde{a}_{ii}}{(1-\tilde{a}_{ii})^3} - \lambda_i + \lambda_{n+i}, \qquad i = 1,2,\cdots n;$$
$$f_{n+1} = \sum_{i=1}^{n} \frac{1}{1-\tilde{a}_{ii}} - \mu;$$

$$f_{n+2} = \sum_{i=1}^{n} \frac{\tilde{a}_{ii}}{(1-\tilde{a}_{ii})^2} - \sigma^2;$$

$$\begin{aligned} f_{n+2+k} &= 2\lambda_k x_k, && k = 1,2,\cdots,2n; \\ f_{3n+2+k} &= x_k^2 + \varepsilon - \tilde{a}_{kk}, && k = 1,2,\cdots,n; \\ f_{4n+2+k} &= x_k^2 + \varepsilon + \tilde{a}_{k-n,k-n} - 1, && k = n+1,\cdots,2n. \end{aligned}$$

$\boldsymbol{C}$ is a (symmetrical) matrix formed from further partial derivatives of f, given as

$$\boldsymbol{C} = \begin{pmatrix} \boldsymbol{U} & \boldsymbol{V} & \boldsymbol{W} & . & . & \boldsymbol{X} & \boldsymbol{Y} \\ \boldsymbol{V}^\tau & . & . & . & . & . & . \\ \boldsymbol{W}^\tau & . & . & . & . & . & . \\ . & . & . & \boldsymbol{P} & . & \boldsymbol{R} & . \\ . & . & . & . & \boldsymbol{Q} & . & \boldsymbol{S} \\ \boldsymbol{X} & . & . & \boldsymbol{R} & . & . & . \\ \boldsymbol{Y} & . & . & . & \boldsymbol{S} & . & . \end{pmatrix},$$

where all the dots denote zero sub-matrices. The non-zero sub-matrices are specified, respectively, as

$$\boldsymbol{U}_{(n\times n)} = \operatorname{diag}\left\{ -\frac{1}{\tilde{a}_{ii}^2} D(i,i) - \frac{1}{(1-\tilde{a}_{ii})^2} D(i,i+1) + \frac{2\theta_1}{(1-\tilde{a}_{ii})^3} + 2\theta_2 \frac{2-\tilde{a}_{ii}}{(1-\tilde{a}_{ii})^4} \right\}_{i=1,2,\cdots,n};$$

$$\boldsymbol{V}_{(n\times 1)} = \left\{ \frac{1}{(1-\tilde{a}_{11})^2} \quad \frac{1}{(1-\tilde{a}_{22})^2} \quad \cdots \quad \frac{1}{(1-\tilde{a}_{nn})^2} \right\}^\tau;$$

$$\boldsymbol{W}_{(n\times 1)} = \left\{ \frac{1+\tilde{a}_{11}}{(1-\tilde{a}_{11})^3} \quad \frac{1+\tilde{a}_{22}}{(1-\tilde{a}_{22})^3} \quad \cdots \quad \frac{1+\tilde{a}_{nn}}{(1-\tilde{a}_{nn})^3} \right\}^\tau;$$

$$\begin{aligned} \boldsymbol{X}_{(n\times n)} &= -\boldsymbol{I}_n; & \boldsymbol{Y}_{(n\times n)} &= \boldsymbol{I}_n; \\ \boldsymbol{P}_{(n\times n)} &= \operatorname{diag}\{2\lambda_k\}_{k=1,2,\cdots,n}; & \boldsymbol{Q}_{(n\times n)} &= \operatorname{diag}\{2\lambda_k\}_{k=n+1,\cdots,2n}; \\ \boldsymbol{R}_{(n\times n)} &= \operatorname{diag}\{2x_k\}_{k=1,2,\cdots,n}; & \boldsymbol{S}_{(n\times n)} &= \operatorname{diag}\{2x_k\}_{k=n+1,\cdots,2n}, \end{aligned}$$

where $\boldsymbol{I}$ is an identity matrix, and the subscripts between brackets of the sub-matrices denote their dimensions.

Appendix 3. Numerical search and initial points

For convenience of analysis we still use u as in (8). When $n = 2^*$, the space is reduced to a plane and the solution intersection for (9) given a pair (μ, σ^2) is reduced to the intersection between a 2-dimensional circle and a straight line, resulting in at most 2 points. This is logical since 2 equations for 2 variables will leave no freedom for relaxed solutions. It is easy to obtain the analytical solution:

$$u_{1,2} = \frac{1}{2}\left(\mu \pm \sqrt{2\sigma^2 + 2\mu - \mu^2}\right).$$

Although a fixed solution can satisfy (9), there is little chance that this is coincidentally the solution for the whole ML equations (7). This implies that in practice, if for some HMM the smallest integer n within the region of (11) is really 2, one should take some value of $n > 2$ in order to let the searching procedure find solutions for the entire (7).

---

* We do not consider the case for $n = 1$ because it gives only a geometrical durational pdf.

In the following discussion, we will assume $n \geq 3$. From $u_i > 1$ and (9) it follows that we should have $\mu > n$. Then if in the data for some HMM $\mu < 3$, we have to modify it to some value $\mu > 3$ before the whole procedure.

For the numerical search not being trapped into some bad point, we need to give some number of initial points and start searching from all these points. From (9) it is clear that permuting the components of $\boldsymbol{u}$ makes no difference for the durational constraints, but it makes a difference for the first equation in (7) which includes also the distribution of acoustic observations. When only one component in $\boldsymbol{u}$ is different while all other components are the same, we get only n initial points by permuting the components. We consider n points as insufficient and design some more points as follows. We take all the $n-2$ components of $\boldsymbol{u}$ to be equal, and another one to have a small difference:

$$u_3 = u_4 = \cdots = u_n = u_2 + \delta, \tag{14}$$

with $\delta > 0$. Then we find the last component on the intersection circle given these $n-1$ values. Putting this into the first equation of (9) we get

$$(n-2)u_n + (u_n - \delta) + u_1 = \mu.$$

From this we solve

$$u_n = \frac{\mu - u_1 + \delta}{n-1}. \tag{15}$$

Putting this and the u_{n-1} from (14) into the second equation of (9) we then get

$$(n-2)\left(\frac{\mu - u_1 + \delta}{n-1} - \frac{1}{2}\right)^2 + \left(\frac{\mu - u_1 + \delta}{n-1} - \delta - \frac{1}{2}\right)^2 + \left(u_1 - \frac{1}{2}\right)^2 = \sigma^2 + \frac{n}{4}.$$

Solving for u_1 and taking arbitrarily the higher value for convenience, we get:

$$u_1 = \frac{1}{n}\left[\mu + \sqrt{(\sigma^2 n + \mu n - \mu^2)(n-1) - \delta^2 n(n-2)}\right]. \tag{16}$$

The above is only one initial point. Since 2 components have different values while all the others are the same, permuting these components will give us $n(n-1)$ initial points (Another $n(n-1)$ points by taking a negative sign before the square root are not used).

The remaining problem is how to choose the value of δ. The condition is to guarantee all $u_i > 1$. This affects the smallest component u_2 most. Using (14) and (15) this gives

$$u_2 = u_n - \delta = \frac{\mu - u_1 + \delta}{n-1} - \delta > 1.$$

To find δ_{max} of δ we combine this with (16) to eliminate u_1 and it follows

$$\left[\delta_{max} n(n-2) + n(n-1) + (1-n)\mu\right]^2 = (\sigma^2 n + \mu n - \mu^2)(n-1) - \delta_{max}^2 n(n-2).$$

Solving this and taking the smaller (safer) value, this gives

$$\delta_{max} = \frac{1}{n-1}\left[(\mu - n) - \sqrt{\frac{\sigma^2(n-1) - (\mu - n)(\mu - 1)}{n-2}}\right].$$

In practice we take some smaller value $\delta < \delta_{max}$ for getting the initial values $\{u_i\}$.

This δ_{max} is only meaningful if the argument of the square root is non-negative, and this casts another lower limit on n for a given σ^2:

$$\tilde{n}_{min} = \frac{\mu(\mu-1)+\sigma^2}{\sigma^2+\mu-1}.$$

Comparing this with (10) we have $\tilde{n}_{min} > n_{min}$, which means that in practice we have to take an $n > \tilde{n}_{min}$ in choosing n. (Recall that n_{min} refers to the case with equal u. This means then that in order to be able to use the initial points chosen this way, it is no more allowed to have equal selfloop probabilities).

On the higher border of n, the data statistics pair (μ, σ^2) of some HMM may not allow any $n < n^L_{max}$ nor even $n < n^U_{max}$. Therefore a reasonable compromise is to relax on one of the two statistics, and preferably on σ^2. Then we need to know the possible range within which σ^2 is allowed to vary, based on the given μ and a chosen n. The procedure of obtaining the range is similar as above but more lengthy. We only give here the resulting range. For the case $n > 3$:

$$\frac{(\mu-n)(\mu-1)}{n-1} < \sigma^2 < \frac{(\mu-n)(\mu-n+2)}{2}.$$

For the case $n = 3$, it requires that n^U_{max} instead of n^L_{max} should be used, namely the solutions of u are located only on the n disjoint arcs. The range obtained is

$$\frac{(\mu-3)(\mu-1)}{2} < \sigma^2 < (\mu-3)(\mu-2).$$

Acknowledgement

The author is very grateful to Louis ten Bosch and Louis Pols for their inspiration and guidance during the research project, and for their comments on this paper.

References

Hochberg, M.M. & Silverman, H.F. (1993): "Constraining the duration variance in HMM-based connected-speech recognition", *Proceedings EUROSPEECH '93, Berlin, Germany, September 1993,* 323-326.

Kamp, Y. (1991): *An introduction to the Baum and EM algorithms for maximum likelihood estimation,* Rapport no. 830 of Institute for Perception Research, Eindhoven.

Lamel, L.F. & Gauvain, J.-L. (1993a): "Identifying non-linguistic speech features", *Proceedings EUROSPEECH '93, Berlin, Germany, September 1993,* 23-30.

Lamel, L.F. & Gauvain, J.-L. (1993b): "High performance speaker-independent phone recognition using CDHMM", *Proceedings EUROSPEECH '93, Berlin, Germany, September 1993,* 121-124.

Lee, K.F. (1989): *Automatic speech recognition: the development of the Sphinx system*, Kluwer Academic Publishers, Boston, Dordrecht, London.

Levinson, S.E. (1986): "Continuously variable duration hidden Markov models for automatic speech recognition", *Computer Speech and Language* **1**: 29-45.

Lloyd, E. (1980): *Handbook of applied mathematics Vol. 2: Probability*, John Wiley & Sons, Ltd. 382-385.

Papoulis, A. (1990): *Probability & statistics*, Prentice-Hall, Inc., Englewood Cliffs, New Jersey.

Press, W.H., Flannery, B.P., Teukolsky, S.A. & Vetterling, W.T. (1989): *Numerical recipes in Pascal*, Cambridge Univ. Press, 305-306.

Rabiner, L.R. & Schafer, R.W. (1978): *Digital processing of speech signals*, Prentice-Hall, Inc., Englewood Cliffs, New Jersey.

Wang, X. (1993a): "Durational modelling in HMM-based speech recognition: towards a justified measure", *Proceedings NATO Advanced Study Institute on New Advances and Trends in Speech Recognition and Coding, Bubión (Granada), Spain, June-July 1993,* Contributed paper, 67-70.

Wang, X. (1993b): "Modelling duration and other long-term speech features in HMM-based speech recognition", *Proceedings of the Institute of Phonetic Sciences of the University of Amsterdam* **17**: 19-32.

Zue, V., Seneff, S. & Glass, J. (1990): "Speech database development at MIT: TIMIT and beyond", *Speech Communication* **9**, 351-356.