Institute of Phonetic Sciences,
University of Amsterdam,
Proceedings 17 (1993), 19-32

MODELLING DURATION AND OTHER LONG-TERM SPEECH FEATURES IN *HMM*-BASED SPEECH RECOGNITION

Xue Wang

Abstract

HMM basically models the speech signals at the level of short-term analysis frames. However, speech information is conveyed not only in this way, but also in long-term features such as the segmental duration. Integration of phonetic knowledge about duration into a recogniser should improve the accuracy of the modelling and therefore increase the recognition performance. In order to justify such approaches, many aspects ranging from philosophical to implementational ones should be dealt with. In this paper a review and discussion is given.

Introduction

Since the introduction of the Hidden Markov Models (HMM) into the engineering applications of automatic speech recognition (ASR), the success is inevitable. Actually it is the HMM that brings forth the first feasible possibility to ASR after many decades of the attempt. The combination of the power of modern computers and the efficient algorithms makes the current encouraging situation possible (e.g. Rabiner et al., 1993).

Standard HMM with the extremely simplified mathematical assumptions already find sufficiently satisfactory applications for moderate tasks. Therefore, based on such standard HMM structure, it is still possible to introduce further refinement for possible improvement and for more difficult recognition tasks. For different application purposes, the direction chosen for refinement may be different. It is also possible to define such directions from a pure mathematical initiative and interests.

The research area can be characterised in other ways as well. Since speech science itself has a much longer history than modern ASR, the well-established theories and rich knowledge can find their position on the new stage. These classical areas include phonetics, phonology, linguistics and psychoacoustics. Scientists in these areas are motivated to inject their knowledge into the ASR, which by itself has been rather engineering from the beginning, therefore can in some respect be quite naïve with regard to the nature of the speech signal that is actually dealt with. On the other hand, where the classical theories find themselves confusing or insufficiently sophisticated, the practice of ASR may also provide new insights to re-justify and enrich the theory.

One observed phenomenon in speech is the varying duration of the phonetically defined segments. The duration has relations with, among other aspects, the phonetic identity and the contexts of the segments, and with the overall speaking rate and style of the utterance. Human listeners use durational cues in the perception process. It is then logical to include a duration mechanism in ASR as well, such that the natural process of speech perception by humans can be better imitated with an ASR. However, the duration is not part of the basic design of the standard HMM. Therefore, it is important to investigate how durational modelling has been implemented in an HMM-based recogniser, and how the durational modelling should affect the recognition performance. Note here that the duration feature is not part of the objective of the recognition (the recogniser does not have to identify the segmental duration, but only the content in an utterance in terms of a sequence of linguistic units). A recogniser can output the segmental duration as a by-product, but it is usually not used. However, a better durational modelling in a recogniser may improve the recognition, provided that the durational feature carries extra information and this information can be made use of by the recogniser.

In this paper, an in-depth review will be given on durational modelling in HMM-based recognition. Attempts will be made to analyse each approach with respect to its impact on the HMM system and possible improvement of the performance. Some general confusions in the understanding of the problem and approaches are clarified.

1. The Durational pdf as a Secondary Measure of HMM

HMM is currently the basis of the most successful technique of automatic speech recognition (Rabiner et al., 1993). This technique is based on statistical methods as opposed to non-statistical rule-based systems. When the rules for speech get more and more complicated while the modelling quality does not get much better, a statistical approach relying less on an understanding of the speech signal wins the battle. However, in essence, the physical reality of speech is neither governed by sets of artificially derived rules, nor generated from a stochastic information source with parameters to be estimated. Statistical modelling based on HMM is only a good approximation to the reality of speech, and therefore many assumptions and compromises have to be made.

In the application of HMM to speech recognition, the speech signal is usually represented as a sequence of feature vectors[1] at discrete time intervals: $O = o_1 o_2 \cdots o_T$. Basically, HMM is used in a recogniser to calculate the probability that O is generated from an HMM. The basic parameters of HMM, namely a in the transition matrix A and b in the observation matrix B, are both defined at a single time step:

$$\begin{aligned} a_{ij} &= P(s_t = j | s_{t-1} = i); \\ b_j(o_t) &= P(o_t | s_t = j), \end{aligned} \tag{1}$$

where P refers to probability, and s is one of the finite states of the HMM.

The durational behaviour of an HMM is usually characterised by a durational probability density function (pdf)[2]. It is defined on all possible durations, and its value is the probability of staying in a state or in an HMM for a given time steps d. This pdf is then

[1] In this study, the actual composition of a feature vector obtained from an analysis frame is not dealt with. The reader can find reference in e.g. Wang et al. (1993) for that aspect. For systems using also time- or frequency-derivatives, we refer to all the components as one vector. Therefore, feature vector, acoustic vector, or observation vector all refer to the same thing, as an analysis frame does.

[2] Note that here and elsewhere if not specified, pdf, mean and variance all refer to the durational distribution, instead of the acoustic distribution.

not a basic quantity of an HMM, namely not a probability of a basic event associated with an HMM. However, the event associated with the durational pdf is a joint event of the basic events. In this sense, the durational pdf is a secondary measure of the process. The durational pdf is defined on a sequence of time frames so it is a long-term feature. Given the Markovain assumption, the durational pdf for a single state i is

$$P(d) = a_{ii}^{d-1}(1-a_{ii}). \tag{2}$$

It can be seen that this is a geometrically decaying function of d. It has been claimed that this is a source of inaccurate modelling with the HMMs since no actual physical events in speech obey this rule.

2. Explicit State Durational Modelling

Owing to the inaccurate durational modelling of the standard HMM at the state level, a very common attempt has been to replace the intrinsic durational behaviour of HMMs with an explicit model at the state level. Actually this explicit state durational modelling is almost taken for granted as the only durational modelling technique. Sufficient work has been done along this direction with concrete conclusions (e.g. Levinson, 1986; Guédon, 1992), therefore it will not be dealt with in detail in this paper.

The principle of the technique is straightforward, namely to replace the durational pdf of each state of a standard HMM with an artificial pdf. The form of this pdf is chosen to be some well established pdf's, such as a Gamma, a multi-nominal, or a Lognormal. The parameters of this pdf are estimated directly from a set of training data. Since the standard HMM formalism is altered, e.g. the transitional behaviour does not follow the Markov process entirely, such a model is called Hidden Semi-Markov model (HSMM). The estimation algorithm for parameters of HSMM (different from HMM) and the explicit pdf can be found in Levinson (1986).

With the parametrical modelling of the durational pdf at the state level, certainly the durational behaviour at this level can be modelled very accurately. All the well chosen pdf functions are governed by two or more parameters, therefore both the mean and variance can be fitted to the statistical distribution of the real data. The HSMM provides some improvement in recognition performance for some tasks (Russell et al., 1985), as compared with the HMM, at the cost of an increased complexity of the whole system.

3. Whole-Model Durational pdf

Although a better durational behaviour at the state level is important, this is not sufficient. The critique of the standard HMM for its state durational behaviour is weakened for systems using HMMs with multiple states. Actually the latter is the most common practice because one wants to model the intra-segment variations of speech as well, with a single model. Within such a model, each state does not necessarily correspond to any well-defined acoustic event with a known durational distribution. In this sense, the durational behaviour of an internal state is less important than that of the whole model, which always corresponds to an acoustic segment. Therefore it is sensible to investigate the whole model durational pdf. If the investigation shows that such a pdf does not behave well enough, it will be then necessary to manipulate it.

In this section, the expression of the whole model durational pdf is derived, starting with an often used simple linear topology as an example. Conclusions for general model topologies are also given with extensions from this simple example. The investigation into the whole-model pdf has only been brought to attention recently (e.g. Guédon, 1992), while the derivation used here is different from his.

The key difference to be taken into account between single state and a number of connected states in the whole model is that the total time d can be spent in different state transitions in many different possible ways. We start from a linear transition topology in which each state can only transit to itself (the selfloop) or to the next state in the cascade. Recall that with the single state the durational pdf decreases with d because the value of the pdf for a larger d is obtained simply by more multiplication with a which are smaller than 1. For a cascade of n state each with a selfloop, the actual way of distributing all the $d \geq n$ time steps[3], namely how many steps to spend in each state, or more generally the state sequence $S = s_1 s_2 \cdots s_T$, is unknown. In this situation, because we are talking about the probability that the process stays within the model for a given d, this value of the pdf is the sum of the probabilities of all possible events, namely all different ways of distributing the total d. This is why in the resulting durational pdf, there can be regions where probability increases with d because these values are sums of many values. Each of the summation terms decreases with d, but the number of terms increases with d. Therefore, when for some regions of d the second effect compensates the first one, the value of pdf can increase with d.

The example is further simplified for situations where all the selfloop probabilities are equal $a_{ii} = a$. It can be seen that the transition probability from any state to its next state are also equal, being $(1-a)$. Then, irrespective of which particular S is taken, the probabilities for taking any single S with a given d are equal. This can be understood in such a way that all the states are mutually indistinguishable with respect to their contribution to the probability of S. It follows that what we need is to calculate the probability of such a single S and the total number of such possible S. (Note here that all the S are different but their *contributions* are equal[4]). The summation over all S is simplified to a multiplication of the probability of a single S with the number of S.

Since each transition from a state to its next takes one step, any sequence S with d will spend exactly n steps in a model with n selfloops, and spend the rest $(d-n)$ to some of the selfloops in an arbitrary way. Such a joint event (for any S) consists of n to-next transitions and $(d-n)$ selfloop transitions. Each of the former has a probability $(1-a)$ and each of the latter, a. Then the probability of the joint event is a multiplication of all these: $a^{d-n}(1-a)^n$. The number of ways of distributing $(d-n)$ time steps (think of them as identical tokens) among the n selfloops (as distinct locations), given by the combinatory formula, is

$$K(d,n) = \binom{d-1}{n-1} = \frac{1}{(n-1)!}(d-1)(d-2)\cdots(d-n+1). \tag{3}$$

Therefore the durational pdf of the model is

$$P_n(d) = K(d,n)a^{d-n}(1-a)^n. \tag{4}$$

[3] It is impossible to have $d < n$ time steps with such a model, therefore the pdf value for those points are simply zero.

[4] For example two distinct sequences S_1 and S_2, with S_1 spending at s_i for 2 and at s_{i+1} for 3, whereas S_2 staying at s_i for 3 and at s_{i+1} for 2. In both cases, this total contribution from s_i and s_{i+1} are equal, being $a^5(1-a)^2$.

This is a negative-binomial distribution (Lloyd, 1980). This discrete distribution is also a close approximation to a continuous Gamma distribution. Such a distribution has a single peak at $d_0 \geq n$ (see Appendix), and has an asymmetric long tail for $d > d_0$. This is ideal for modelling segments such as the phones.

For the more general situation with different selfloop probabilities, we start with two states in cascade. We still denote the total time spent in these two states by *d*. It is obvious that the more time is spent in state *i*, the less time is left for state *i*+1, with the total always being *d*. For each particular number of time steps in *i*, we have exactly one way of distribution. Therefore the summation over all possible ways is simply a summation over all possible values of times spent in the first state *i*, namely

$$P_2(d) = \sum_{l=1}^{d-2} a_{ii}^{l} a_{i,i+1} a_{i+1,i+1}^{d-2-l} a_{i+1,i+2}. \tag{5}$$

Note that we also considered the time spent to transit out of each state.

It can be seen that this formula is formally a convolution[5] between *a* terms. However this is obtained by distributing the time steps among states, therefore this has nothing to do with the convolution used in e.g. calculating the output from a digital filter. If an explicit durational term is used to replace each a_{ii} as is done in HSMM, this convolution formula still holds. This can be directly extended to a longer cascade with more than two states by convoluting gradually the resulting cascades with other ones, with some minor modification for the initial value of *d*. For a model topology with arbitrary transition branches, as long as there are no large feedback loops involving more than one state (namely if the whole model is still left-to-right, the only type used to model speech), the total durational pdf is simply a contributed sum from all linear branches weighted by the probabilities of taking these branches.

It is obvious that the simpler linear case with equal *a* is a special case of the general cases for single linear branch with different a_{ii}, namely the simple one can also be obtained with a convolution. But the result of the convolution for the case of equal *a* is known with a closed form, being a negative-binomial function. It can be speculated from this that the convolution with different and more than two a_{ii} will result in similar functions. It is difficult to obtain a closed form of it, but a numerical procedure for calculating it can be derived. However, for the following reasons, we do not go further along this line. For the purpose of fitting the durational behaviour of a whole model, even a binomial case with the simplest topology and equal *a*, as shown above, may be competent. The more complicated cases including different a_{ii}, parallel transitional branches, and eventually the convolution with the parametrical terms of HSMM[6], do not seem necessary, at least before we have tried the simpler setups. Furthermore, within the framework of statistical modelling, any attempt which increases the complexity of the system and the number of parameters which have to be trained with limited data, should be avoided as much as possible.

[5] This derivation is obtained with our special case without using knowledge in probability theory (Lloyd, 1980) which says that the distribution of the sum of independent random variables is a convolution of the distributions of individual variables.

[6] If the parametric term is a Gaussian, the convolution will still result in a Gaussian (with different parameters in general). Gaussian is however not often used as durational pdf because its symmetrical tails are not realistic. Other functions will in general result in complicated forms after convolution.

4. Acoustic Durational pdf

In the above discussion of durational pdf, we have only used the transitional probabilities A, not the observation probabilities B. One argument for this is that the transitional behaviour determines the duration that it models. However, the moment that the process enters a state is also governed by the associated process of observing a particular acoustic vector $\boldsymbol{o}$. In other words, the durational behaviour can also be influenced by the temporal structure of the observation sequences O and B which store statistical information of the training data.

It is important to distinguish which set of speech data is being concerned, being the whole set of training data, or a particular sequence O to be recognised. In this section, it will be clarified that the durational pdf in the previous section is a pdf that has conceptually considered all theoretically possible O, either being present in the training data or not. Therefore it is called a *full* pdf referring to its full consideration. Note that it has been taken for granted in the literature that this full pdf is *the* pdf that characterises the durational behaviour of an HMM, without realising from what statistical information it has been calculated. It requires further investigation to clarify whether a full pdf is a justified measure of the durational behaviour (Wang, 1993).

In the procedures of both training and recognition, it is usually a particular set of data, instead of a full set of data, that play the role in the calculation. The phonetic knowledge that segmental duration affects spectra (e.g. van Son et al., 1990, 1992) also suggests a need for a pdf which is more closely related to the acoustic data. This durational pdf, called *acoustic* pdf, will also be proposed in this section.

The derivation starts with a relation between the full pdf and $P(O|\lambda)$, the quantity very often used and referred to as the likelihood that a model (with parameter set $\lambda = \{A,B\}$) generates O. Similar to the standard forward procedure (see e.g. Juang et al., 1992) where a probability of the observation of a particular partial sequence $o_1 o_2 \cdots o_t$ is defined as

$$\alpha_i(t) = P(o_1 o_2 \cdots o_t, s_t = i|\lambda), \tag{6}$$

we define an observation probability of any partial sequence

$$\tilde{\alpha}_i(t) = P(\forall(o_1 o_2 \cdots o_t), s_t = i|\lambda). \tag{7}$$

The calculation is also similar to that of α, namely in a way of recursion with t, except that an extra summation taken over all possible different o is performed at each time t. It is obvious that this is equivalent to considering all different possible partial sequences as a whole. The recursion is

$$\tilde{\alpha}_j(t) = \sum_{o_t}\left[\sum_{i=1}^{N} \tilde{\alpha}_i(t-1) a_{ij} b_j(o_t)\right]. \tag{8}$$

Since $\tilde{\alpha}_i(t-1)$ has considered all different partial sequences till $(t-1)$ and a is constant, the summation over o_t only applies to b terms, which leads to unity (e.g. for a continuous-observation system),

$$\sum_{o_t} b_j(o_t) = \int_{o_t} b_j(o_t) do_t = 1, \tag{9}$$

where the integral is performed over the entire multi-dimensional acoustic space. It follows that the recursion is only over all different S, using only the A parameters,

$$\tilde{\alpha}_j(t) = \sum_{i=1}^{N} \tilde{\alpha}_i(t-1)a_{ij}. \tag{10}$$

The recursion terminates with the likelihood that λ generates any $\boldsymbol{O}$, each with a total duration d

$$P(\forall(o_1, o_2, \cdots o_d)|\lambda) = P(\forall(\boldsymbol{O}(d))|\lambda) = \sum_{i=1}^{N} \tilde{\alpha}_i(d). \tag{11}$$

In the whole forward recursion procedure, all S have been considered by means of considering at each t all state transitions. The summation of the joint likelihood $P(\boldsymbol{O}, S)$ over all $\boldsymbol{O} \in \{\boldsymbol{O}(d)\}$ results in the marginal $P(S)$, but the duration of all these S are confined to d. A further summation of $P(S)$ over all such S is exactly the same operation as in the previous section, only in the latter calculation, a different way of listing all S was used. Therefore the likelihood in (11) equals $P_n(d)$. Combining these we obtain

$$P_n(d) = P(\forall \boldsymbol{O}(d)|\lambda) = \sum_{\boldsymbol{O}(d)} P(\boldsymbol{O}(d)|\lambda). \tag{12}$$

From this expression we can give an alternative interpretation of the full pdf as:

the likelihood that a model generates any observation sequence with a given duration.

The relation (12) between the full pdf and the likelihood indicates that all possible $\boldsymbol{O}$ have been considered. However, many of these $\boldsymbol{O}$ are very unlikely to occur in actual speech data. It can be argued that these unlikely $\boldsymbol{O}$ would contribute very little to the summation even if they would be available, therefore the full-pdf would still provide a correct data statistics. However, it has not been proven whether such a penalisation is sufficient since the training of models follows a standard Baum-Welch algorithm in which the fit of the pdf to data is not used as an optimisation criterion. This can also be seen in the divergence between the full-pdf of a model trained with a data set and the data pdf of the same set, as given by the bars and curves in Fig.1. and 2. The data pdf is from a histogram with the total counts given between brackets.

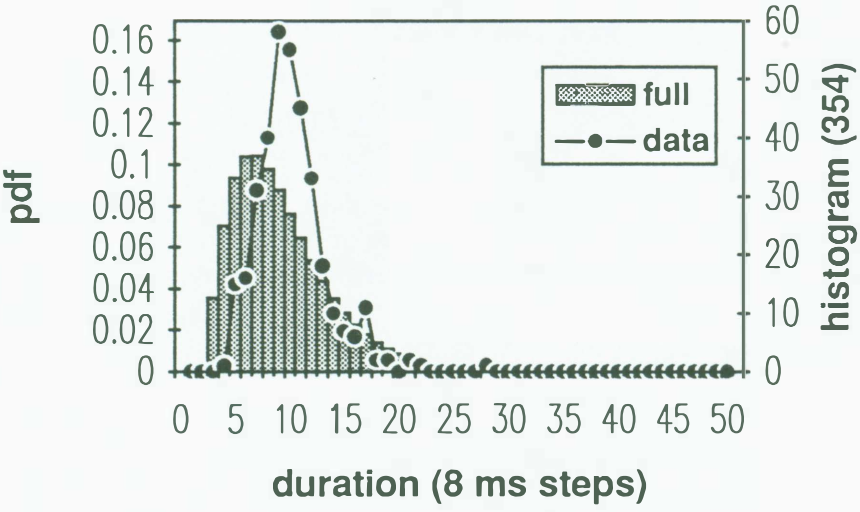

Fig. 1. The durational pdf's of the Dutch vowel /ɑ/.

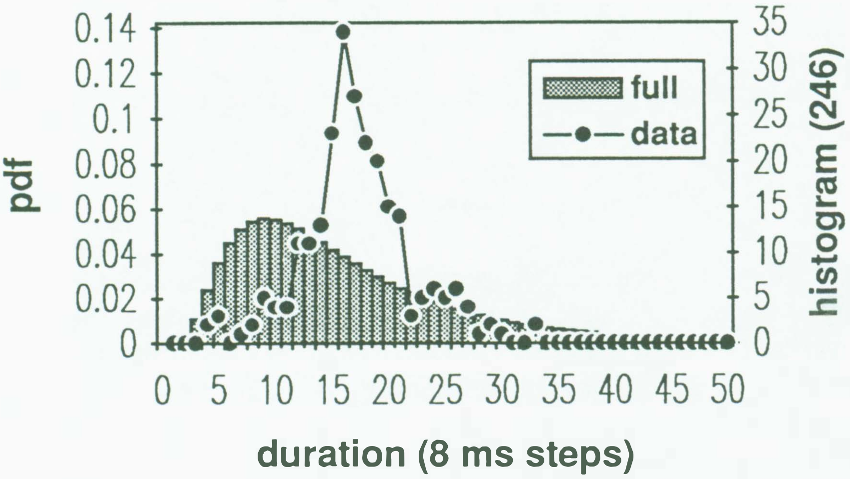

Fig. 2. The durational pdf's of the Dutch vowel /a/.

It can be seen that one possible way to get an acoustic pdf is to use only those available O within each subset $\{O(d)\}$ of the data with duration d, in the summation of (12). However, direct substitution gives a problem. The value of the likelihood $P(O|\lambda)$ decreases geometrically with duration d of O due to multiplication of probabilities. In the calculation of the full pdf, this is not a problem since for longer d, the total number of O also increases geometrically, so the summation over this larger number of O will weight it back. But with only the available set, this is the problem since the number of O in $\{O(d)\}$ is simply provided by the data, and the *a priori* distribution of the numbers is far from geometrically increasing.

In order to use the available O to estimate a correct acoustic pdf, we assume that each O in the data set is a representative of all possible O. Before summing, each likelihood is first down-scaled by the total number N_d of O in $\{O(d)\}$, and then up-scaled by an assumed number of all possible O, given by a discrete-type observation. This estimated measure is normalised through the total range D of duration, and a proposed calculation is then

$$\hat{P}_n(d) = \frac{\left(\frac{M^d}{N_d}\right) \sum_{O\in\{O(d)\}} P(O|\lambda)}{\sum_{d=1}^{D} \left(\frac{M^d}{N_d}\right) \sum_{O\in\{O(d)\}} P(O|\lambda)}. \qquad (13)$$

In this formula, M is an assumed codebook size. This and other proposed formulae for acoustic pdf can only be useful if the limited size of data do not introduce unacceptable irregularities. This is a serious practical problem since a whole set of data with a moderate size will be further partitioned for different durations, leaving the size of $\{O(d)\}$ unacceptably small.

5. Constraint on the Whole-Model Duration

Being aware of the importance of the whole-model pdf instead of merely adjusting the single-state pdf, a step forward is to actively fit the whole-model pdf to the empirical durational distribution of the data. The only known activity along this line is reported in a very recent paper (Hochberg et al., 1993), using whole-word models in a connected alphadigit task. The actual approach is to use Gaussian HSMM to fit the state mean and variance, and to constrain the state variances on the word variance. Improvement in recognition (10.9% error rate as compared to 12.3% without constraint on word duration) has been achieved even with this simple choice of Gaussian pdf.

This approach bypasses the investigation of the whole-model pdf (and trying to modify it). It directly goes to the estimation of the state parameters. In the whole procedure of estimation with the constraint, it is a particular state sequence S, instead of all possible S, that is considered. This technically replaces the calculation of convolution by a much simpler multiplication. When an S is known, namely a certain number of time steps is devoted to each state (of a linear model again), the probability of taking S is obtained by multiplying the parametrical HSMM terms of all states, and the transition probabilities between states.

Both the lower-order statistics (mean and variance) of whole model and of single state have simple relations, such that the former are sums of the latter. The approach is still to estimate the values of the state variance of the explicit pdf terms, while these values are further constrained by the whole-model variance, the latter being obtained empirically from data. This way, although the variance of each single state is not fitted to the data, the variance of the whole model is. It turns out that the actual estimation of the variances does not affect the estimation of the mean of each individual state, so that the latter can still be fitted to the data.

Although both convolution and multiplication coincidentally result in Gaussian pdf's, the parameters of them are different (usually the convolution results in wider distribution, namely lager variance, than the multiplication does). The investigation here motivates us to have a closer look at the whole-model full-pdf in section 3. Although the full-pdf of a standard HMM has a good form of pdf, its parameters need to be adjusted with algorithms in addition to the standard training.

It is interesting to view this phenomenon from a different perspective, namely in combination with the acoustic pdf. Although the coincidence that both convolution and multiplication of Gaussian pdf result in Gaussian, which only applies with this special function, the two resulting Gaussian pdf's represent conceptually different kinds of probability. When we start with the general joint probability (we drop the parameter indication λ because it is not relevant here)

$$P(\boldsymbol{O},S) = P(\boldsymbol{O}|S)P(S), \tag{14}$$

the result from the multiplication is simply the $P(S)$ as the last term, namely the probability of a particular S. With our induction, the full pdf is

$$P(d) = \sum_{\boldsymbol{O}\in\{\boldsymbol{O}(d)\}} P(\boldsymbol{O}) = \sum_{\boldsymbol{O}\in\{\boldsymbol{O}(d)\}} \sum_{S} P(\boldsymbol{O},S). \tag{15}$$

This also equals a convolution of the individual state pdf, obtained from another way of induction. Our attempt to obtain an acoustic pdf applies to the outermost summation, namely after the summation over S is done. Then we try to choose only those existing $\boldsymbol{O} \in \{\boldsymbol{O}(d)\}$. $P(S)$ essentially lies on one extreme, namely only one S (and associated with only one $\boldsymbol{O}$) is used, whereas the full pdf $P(d)$ lies on the other, i.e. it has summed over all $\boldsymbol{O} \in \{\boldsymbol{O}(d)\}$ and all S associated with them.

One attempt to obtain an acoustic pdf, therefore, can be to sum over only a selected set of S associated with each particular $\boldsymbol{O}$, or simply one S for each $\boldsymbol{O}$, found by a Viterbi scoring, since that S is dominating anyway. By doing so, the acoustic pdf we get should be closer to $P(S)$ than the acoustic pdf using all S. Using Gaussian state pdf, since both extremes are Gaussian, the acoustic pdf obtained this way should also be a Gaussian.

Looking at the lower-order statistics is a parametrical way of statistical modelling. In one situation, either the (sufficient) statistics, or the whole pdf, will be more convenient and insightful. The inaccuracy left by merely using the statistics can be treated in a later step. In this light, looking at the first two statistics mean and variance, either with or without using a parametric model, may be an alternative way to looking at the entire distribution pdf.

6. Duration Relative to Speaking Rate

Speaking rate is a global measure of an utterance, and can be defined as e.g. number of segments per unit of time. It has a certain relation with the segment duration. It is another way to see the same problem as treated with durational modelling directly. Due to the difference in the formalism in the two approaches, the actual kinds of measures, the emphasis on the aspects of the variations in the data, and eventually the performance improvement they will bring, may be different.

Jones et al. (1993) in a recent work tried to incorporate rate information into large vocabulary continuous speech recognition. The essence of this work is to attach to the set of HMMs some parallel models in which to store direct durational statistical information of the data. These parallel models differ from the HMM or HSMM in that they are not part of the modelled stochastic process that generate the speech. The statistical information stored in them are used to alter the scores of the recognition hypotheses when the speaking rate is available.

The speaking rate of an unknown utterance is not available when the utterance enters the recogniser. Some procedure is needed to get it or to estimate it. Actually in this work the whole recognition is almost finished when the rate is estimated. An N-best algorithm is applied that uses the set of basic HMMs (before any adaptation process to the rate) to get a list of N hypotheses. Based on the segmentation from these hypotheses, an estimated rate (or duration) of each basic unit is obtained. This rate information is used to calculate a rate score for each unit. The rate score for the whole utterance combined in a way with the basic score is used to re-score each of the hypotheses. The new top hypothesis provides the recognition.

Three ways of defining the rate models were tested, each with certain advantages. These are (1) partitioned model which splits the data into 3 categories being fastest, average and slowest; (2) shifted-mean model; and (3) relative normalised duration model. The partitioned model, though the simplest one and having the problem of less training data than the other two, gave the best improvement. The actual ways of storing the rate information include using the absolute minimal and maximal duration, the interval between which 90% of the duration occurred, a smoothed durational histogram from the data, and the mean and variance of an assumed Gaussian distribution. The actual ways of using these different types of durational information in the re-scoring process are different, e.g. for the partitioned models this is simply to choose the appropriate partition of the models. The best performance was provided by the histogram, being 10% word error reduction relative to the baseline without rate adaptation.

7. Beyond the HMM: Modelling in the Whole System

There can be a confusion between hidden Markov models and modelling speech in general. It can be seen from the title of this article that these two are not necessarily the same. Obtaining improved, or even perfect, durational behaviour of all the HMMs in a system is not the only way to cope with the durational information, and only doing so does not guarantee a complete incorporation of the information. Review and analyses in section 1 through 5 have been concerned with the HMMs themselves. The approaches in section 6 show the incorporation of durational information into the system, but outside the HMMs.

In the general modelling problem, hidden Markov models are only parts of all models used in a speech recogniser to model the whole speech. This can already be seen from the fact that the language models play a substantial role in recognition. As the duration phenomenon is concerned, it should also be modelled outside the HMMs. Furthermore, the duration itself is present at different levels of the whole structure of speech. It will be improper to just try to model all durational variations within the HMMs, which are still segmental models. A particular serious situation is in systems with sub-word units, e.g. in a continuous speech recogniser. Merely modelling the duration of the basic units (e.g. phones) cannot model the duration of the words composed of phones.

It is general in a statistical approach of recognition to let the system learn as much as possible the information in the training set, including the shape of and distribution within the acoustic space (mainly static statistics), and time correlation between units at various levels ranging from frames to words. Although this statistical approach seems blind in the sense that we let the system learn by itself, we have to, and are able to design the structures for the system to store the statistics. Some of the correlation is embedded in the structure so that they are not to be learned from data while others are to be learned. However, only the values of the parameters, not the structures themselves, will be altered during the training. Therefore a meaningful new approach within the statistical framework should always involve new designs of structures.

8. Duration and Phonetic Contexts

It is a common observation that the segment duration is affected by the phonetic context the segment is in. Therefore it is natural to take this into account in durational modelling. One way to do this is to model the duration of a segment with different models for different contexts. The use of triphones, having a structure with fixed type of context, is one way for coping with duration in context. A similarity can be seen between such an approach and the partitioned model in section 6. However, direct use of a triphone-inventory may not be appropriate since different triphone contexts can have the same durational impact, on the central phone. Furthermore, longer contexts are not considered by triphones, whereas attempt to use m-phones with $m > 3$ certainly brings complexity problem. Some entries of the longer contexts may be better identified with part-of-speech, e.g. the location associated with pre-pausal lengthening.

The results of such investigations will involve new phonetic knowledge. Therefore work along this line will give insights not only in HMM-based recognition, but also in phonetics itself. Because the contextual effect on segment duration is a complicated phenomenon, certain descriptive models are needed to specify these relations (van Santen et al., 1990; van Santen, 1992).

9. Combination of Knowledge and Statistical Approach

There should be no contradictions between a statistical approach of durational modelling and a knowledge-based approach. Both ideas must be actively used.

The segment duration as an observed phenomenon is an aspect of variation in speech. In order to have some measure on this phenomenon, statistics is a good way to cope with the uncertainty. Statistics is also good in dealing with relations between probabilities of different kinds, and in different parts. Although statistics is not the only possible way, it is much more sophisticated and richer than e.g. only indicating the absolute duration of a segment without any probability associated with it. We can compare this with the experiment of throwing dice. Only observing the number on the top side, we can only get 6 values, and we do not have any indication of a test e.g. with more throws. Recording some counts, e.g. the number of two consecutive throws with the total value of 7, will introduce more measures, and can virtually define new events and characterise unlimited number of different tests. Even stochastic processes can be defined. Going back to the duration in speech, some of the events do not even need to be defined: they exist.

Using statistics, however, does not imply that human knowledge cannot enter. One most straightforward way to provide a forum for the knowledge is the process of defining probabilistic events. The probability theory can help us define new events with basic, simple ones, using partition, joining, or conditioning properties. The example operations we have seen above are, to partition with the durational intervals, and to condition with the contexts. Without our knowledge about speech, we cannot define these events and manipulate these probabilities.

An understanding of the problem of statistics and knowledge will pave the basis of the ways how a system equipped with durational knowledge should work. For example, we want our recogniser to use its optimal subset of models to speech with particular speaking rate. Then an external informing signal should be available that tells the system about this rate. For this purpose the aforementioned *N*-best can provide the signal while the duration-partitioned models can be the set to be selected. Note that the partition is defined with our knowledge, whereas within each partition, statistics is needed.

10. Still Open Questions

It is not clear whether the durational modelling is necessary, namely manipulations in other parts of the system, or some implicit modelling simply have done the same job already. It is neither clear, similarly, whether a durational modelling only alters the durational behaviour of a system, nothing else. Due to the complexity of the structured approach of all different durational modelling, and the complexity of the whole recogniser, these questions will only be answered in practice, but not in a theoretical or philosophical way.

Other long-term speech features such as an energy or pitch contour, also find themselves difficult to be integrated in an HMM-based recogniser. Ways of integration should be found, and the mismatches of timing and dynamic range between the long-term features and the frame-based ones, should be considered properly (see e.g. Juang et al., 1992). Techniques found useful in durational modelling, may also be used for any long-term features.

Acknowledgments

The author thanks Louis ten Bosch and Louis Pols for their careful reading of the draft, discussions and useful comments on both the text and the general topics of research. He also thanks Mark Huckvale for useful discussions during the time he spent at University College London, under an ERASMUS exchange scheme. Part of the work in this article was carried out on an RS/6000 computer which was kindly provided by IBM.

Appendix: Full pdf of Standard HMM

In this appendix, we discuss the behaviour of a full pdf of a whole standard HMM, with a linear topology of n selfloops and equal a. The pdf re-written from (4) is

$$P_n(d) = \begin{cases} 0 & , \quad d < n; \\ (1-a)^d & , \quad d = n; \\ \dfrac{1}{(n-1)!}\left(\dfrac{1-a}{a}\right)^n (d-1)(d-2)\cdots(d-n+1)a^d & , \quad d > n. \end{cases} \qquad \text{(a1)}$$

In order to find the maxima of this negative-binomial pdf, we look at the cases for $d > n$ and treat $d = n$ as a special case. For the purpose of analysis, we extrapolate d to a continuous variable. Since a log function is monotonic with its argument, the maxima of a function will have the same locations as its log. For convenience we take the (natural) log of the pdf, also leaving out the constant term,

$$Q(d) = \log(d-1) + \log(d-2) + \cdots + \log(d-n+1) + d\log a. \qquad \text{(a2)}$$

Derivation with respect to d gives

$$Q'(d) = \frac{1}{d-1} + \frac{1}{d-2} + \cdots + \frac{1}{d-n+1} + \log a. \qquad \text{(a3)}$$

Since $a < 1$, the last term $\log a < 0$. Because $d > n$, all the other terms are positive and so is the sum of them. Furthermore, it is obvious that all the terms except the last one is monotonically decreasing with d and therefore the sum of them is so, too. Therefore, there will be at most one point at which the positive and negative terms compensate to make the only critical point $Q'(d_0) = 0$. Since the second derivative is negative everywhere, this critical point is a maximum, and its location d_0 is given by

$$a = \exp\left\{-\left(\frac{1}{d_0-1} + \frac{1}{d_0-2} + \cdots + \frac{1}{d_0-n+1}\right)\right\}. \qquad \text{(a4)}$$

Note that the actual value of d for the maximum of the pdf can only take an integer. If this is not exactly d_0, the true maximum is at one of its two neighbour points:

$$\hat{d}_0 = \arg\max_d \{P_n([d_0]), P_n([d_0]+1)\}, \qquad \text{(a5)}$$

where $[d_0]$ truncates d_0 to its integer part.

Further analysis reveals the following behaviour of the pdf with respect to n, a and d. For a given a, a larger n gives a larger number of positive terms in (a3). Therefore d

has to be larger to make the sum of the positive terms small enough to be compensated with the negative term. This means that a longer model with more selfloops will have its maximum towards larger *d*. On the other hand, *a* should satisfy certain conditions to guarantee that the maximum is located at $d_0 > n$. Putting this into (a4) we then get

$$a > \exp\left\{-\left(1+\frac{1}{2}+\cdots+\frac{1}{n-1}\right)\right\}. \tag{a6}$$

For example, the required value of a for $n=2$ is $a > e^{-1} \cong 0.3679$, and for $n=4$ is $a > e^{-11/6} \cong 0.1599$. If this condition is not satisfied, the maximum is at $d_0 = n$, namely, the pdf peak only has one slope on the right side, and the left side is a sudden 'cliff'.

References

Guédon, Y. (1992): "Review of several stochastic speech unit models", *Computer Speech and Language* **6**: 377-402.

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

Jones, M. & Woodland, P.C. (1993): "Using relative duration in large vocabulary speech recognition", *Proceedings EUROSPEECH '93, Berlin, Germany, September 1993*, 311-314.

Juang, B.-H. & Rabiner, L. R. (1992): "Issues in using hidden Markov models for speech recognition", in: Furui, S. and Soudhi, M. M. (eds.), *Advances in Speech Signal Processing*, Marcel Dekker, Inc. New York, 509-553.

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 applicable mathematics, Volume II: Probability*, John Wiley & Sons Ltd. Chichester.

Rabiner, L. R. & Juang, B.-H. (1993): *Fundamentals of speech recognition*, Prentice-Hall Inc., Englewood Cliffs, New Jersey.

Russell, M.J. & Moore, R.K. (1985): "Explicit modelling of state occupancy in hidden Markov models for automatic speech recognition", *Proceedings ICASSP-85, Tampa, Florida, March 1985*, 5-8.

van Santen, J.P.H. & Olive, J.P. (1990): "The analysis of contextual effects on segmental duration", *Computer Speech and Language*, **4**, 359-390.

van Santen, J.P.H. (1992): "Contextual effects on vowel duration", *Speech Communication*, **11**(6), 513-546.

van Son, R.J.J.H. & Pols, L.C.W. (1990): "Formant frequencies of Dutch vowels in a text, read at normal and fast rate", *J. Acoust. Soc. Am.* **88**(4), 1683-1693.

van Son, R.J.J.H. & Pols, L.C.W. (1992): "Formant movements of Dutch vowels in a text, read at normal and fast rate", *J. Acoust. Soc. Am.* **92**(1), 121-127.

Wang, X. (1993): "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., ten Bosch, L.F.M. & Pols, L.C.W. (1993): "Impact of dimensionality and correlation of observation vectors in HMM-based speech recognition", *Proceedings EUROSPEECH '93, Berlin, Germany, September 1993*, 1583-1586.