Scaled Random Trajectory Segment Models

by

Jacob Goldberger and David Burshtein

Dept. Electrical Engineering – Systems

Tel-Aviv University, Tel-Aviv 69978, Israel

email: jacob@eng.tau.ac.il

Abstract

Speech recognition systems that are based on hidden Markov modeling (HMM), assume that the mean trajectory feature vector within a state is constant over time. In recent years, segment models that attempt to describe the dynamics of the speech signal within a phonetic unit, have been proposed. Some of these models describe the mean trajectory over time as a random process. In this paper we present the concept of a scaled random trajectory segment model, which aims to overcome the modeling problem created by the fact that segment realizations of the same phonetic unit differ in length. The new model is supported by a direct experimental evidence. It offers the following advantages over the standard (non-scaled) model. First, it shows improved performance compared to the non-scaled model. This is demonstrated using phone classification experiments. Second, it yields closed form expressions for the estimated parameters, unlike the previously suggested, non-scaled model, that requires more complicated iterative estimation procedures.

1 Introduction

The standard hidden Markov model (HMM) provides a powerful technique for representing speech utterances by a piecewise stationary process. The model assumes the existence of states, such that the observations are locally independent and identically distributed (IID) within a state. However, empirical evidence indicates that the feature vectors that correspond to some of these states clearly violate the IID assumption. In recent years, alternative models, that attempt to overcome this limitation were proposed and implemented in automatic speech recognition systems (Ostendorf, Digalakis and Kimball, 1996). These methods are usually known by the name segment models, since the fundamental modeling unit is a segment, consisting of a sequence of frames that represent some phonetically meaningful speech unit. A segment is commonly taken to correspond to a phone, but it may, however, correspond to a sub-phonetic event (one phone may be represented as a sequence of several segments). This modeling approach is opposed to the HMM paradigm, that utilizes frame-based modeling.

Various segment models have been proposed, e.g. (Deng, Aksmanovic, Sun and Wu, 1994), (Ghitza and Sondhi, 1993), (Digalakis, Rohlicek and Ostendorf, 1993) and (Goldberger, Burshtein and Franco, 1997). A comprehensive survey on the subject can be found in (Ostendorf, Digalakis and Kimball, 1996). A number of studies have proposed stochastic descriptions of the mean trajectory, as an alternative to the multi description of the mean trajectory, that is provided by an HMM whose state distributions are mixtures of Gaussians. This concept of random trajectory segmental modeling (RTSM) was first suggested by Russell (1993) who named this approach 'segmental HMM'. We prefer to use the more specific term RTSM because it better reflects the particular characteristics of the model that distinguish it from other segmental models. RTSMs can be thought of as a generalization of the Gaussian HMM formalism. The main difference is that the mean of the acoustic feature vector in a state, is not a fixed parameter. Instead, it is a random variable sampled once for each state transition. The acoustic motivation for this framework is

that we wish to separately model two distinct types of variability: long term variations, such as speaker identity, and short term variations which occur within a given state as a result of random fluctuations. The long term variability is modeled by a probability density function (PDF) used to select the sampled mean. The short term variability within a state is modeled by the deviation of the feature vectors from the sampled mean. In standard HMM these two effects are modeled implicitly by a single PDF.

A disadvantage of RTSMs is that segment realizations with varying durations are not properly scaled. In this paper we present the concept of a scaled RTSM. The new model is supported by a direct experimental evidence. It offers the following advantages over the standard (non-scaled) model. First, the new model shows improved performance compared to the non-scaled model. Second, it yields closed form expressions for the estimated parameters, unlike the standard, non-scaled model, that requires more complicated iterative estimation procedures.

In Section 2 we review the RTSM approach. In Section 3 we present the new proposed scaled RTSM and derive recognition and parameter estimation algorithms for the new model. In Section 4 we extend our results to scaled linear RTSMs. In Section 5 we demonstrate the performance of the new proposed models and compare it with standard RTSMs, using a phone classification task. In Section 6 we conclude our results.

2 Static Random Trajectory Segment Models

We first provide a formal description of the random trajectory segmental modeling approach. Let $f_s(\gamma)$ be a PDF defined on some family of valid trajectories, γ, at a given state, s. On arrival at state s, a trajectory is chosen according to this PDF. Once γ is determined, we can model the within-segment variation at each frame independently. Denote by $f_s(x_t|\gamma, t)$ the PDF of the frame x_t given the chosen mean trajectory and the time index t. The PDF of the segment data realization $x = (x_0, ..., x_{n-1})$ is given by

$$P_s(x) = \int_\gamma f_s(\gamma) \prod_t f_s(x_t|\gamma, t) d\gamma$$

According to Russell's terminology (Russell, 1993), $f_s(\gamma)$ accounts for extra-segmental variation which would lead to different trajectories for the same phonetic unit, while $f_s(x_t|\gamma, t)$ hopefully accounts for much smaller intra-segmental variations in the realization of a particular trajectory.

The PDF $f_s(\gamma)$ can be either discrete or continuous. In the rest of the section we concentrate on the simplest continuous case where the trajectories PDF is Gaussian and the mean trajectory is constant over time. This static RTSM was originally presented by Russell (1993). More precisely, a static RTSM assumes that the observations within a state $x = (x_0, ..., x_{n-1})$ are generated according to

$$x_t = \mu + a + \epsilon_t \qquad t = 0, ..., n-1 \tag{1}$$

where μ is a fixed parameter, associated with the state, that describes the grand mean trajectory. The random variable a is a shift of the mean trajectory that is global to the entire segment realization. It is assumed that $a \sim N(0, \sigma_a^2)$ (i.e., a is a Gaussian random variable with mean 0 and variance σ_a^2). The short term variability is represented by ϵ_t, which is a zero mean Gaussian random variable with state dependent variance, $\epsilon_t \sim N(0, \sigma^2)$. To simplify notation, it will be assumed that the observations are one dimensional. Generalization to the multi-dimensional case is straight-forward. The PDF of the segment data realization, x, is given by

$$f(x) = \int_a f(x, a) da \tag{2}$$

$$f(x, a) = f(a) \prod_t f(x_t|a) \tag{3}$$

where

$$f(a) = \frac{1}{\sqrt{2\pi}\sigma_a} e^{-\frac{a^2}{2\sigma_a^2}} \qquad f(x_t|a) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2\sigma^2}(x_t - \mu - a)^2} \tag{4}$$

In this presentation of the static RTSM, the probability expressions are formulated in terms of the shift of the trajectory from the model mean, rather than the location of the trajectory itself as in (Russell, 1993) and (Gales and Young, 1993). We are adopting this convention because it explicitly reflects the linear behavior of the model (equation 1).

Two methods for computing the likelihood score have been proposed. The first one (Russell, 1993) is a maximum a posteriori (MAP) approach. The MAP method approximates $f(x)$ by $f(x, \hat{a})$, where $\hat{a}$ is the shift which maximizes the joint PDF $f(x, a)$. That is to say, segment identification is made based on the segment hypothesis for which $f(x, \hat{a})$ is maximized. A closed form for $\hat{a}$ is obtained by setting the derivative, $\partial \log f(x, a)/\partial a$ equal to zero, thus yielding,

$$-\frac{a}{\sigma_a^2} + \frac{1}{\sigma^2}\sum_{t=0}^{n-1}(x_t - \mu - a) = 0$$

Hence,

$$\hat{a} = \arg\max_a f(x, a) = \frac{\frac{1}{\sigma^2}\sum_{t=0}^{n-1}(x_t - \mu)}{\frac{1}{\sigma_a^2} + \frac{n}{\sigma^2}} \tag{5}$$

We use the following notation

$$E_x = \frac{1}{n}\sum_{t=0}^{n-1} x_t \quad , \quad V_x = \frac{1}{n}\sum_{t=0}^{n-1} x_t^2 - E_x^2 \quad , \quad V_n = \left(\frac{1}{\sigma_a^2} + \frac{n}{\sigma^2}\right) \tag{6}$$

Using this notation, we have,

$$\hat{a} = \frac{n}{\sigma^2 V_n}(E_x - \mu) \tag{7}$$

The last expression will be used to help evaluate $f(x)$.

Following Gales and Young (1993), we now derive a closed form expression for $f(x)$. By (3) and (4),

$$f(x, a) = \frac{1}{\sqrt{2\pi}\sigma_a}\left(\frac{1}{\sqrt{2\pi}\sigma}\right)^n e^{-\frac{1}{2}g(x,a)} \tag{8}$$

where

$$\begin{aligned} g(x, a) &= \frac{a^2}{\sigma_a^2} + \frac{1}{\sigma^2}\sum_t (x_t - \mu - a)^2 \\ &= \left(\frac{1}{\sigma_a^2} + \frac{n}{\sigma^2}\right) a^2 - 2a\frac{1}{\sigma^2}\sum(x_t - \mu) + \frac{1}{\sigma^2}\sum(x_t - \mu)^2 \end{aligned}$$

$$
\begin{aligned}
&= V_n a^2 - 2a\frac{n}{\sigma^2}(E_x - \mu) + \frac{1}{\sigma^2}\sum_t (x_t - \mu)^2 \\
&= V_n(a - \frac{n}{\sigma^2 V_n}(E_x - \mu))^2 + \frac{1}{\sigma^2}\sum_t (x_t - \mu)^2 - \frac{n^2}{\sigma^4 V_n}(E_x - \mu)^2 \\
&= V_n(a - \hat{a})^2 + \frac{n}{\sigma^2}\left(V_x + \frac{1}{\sigma_a^2 V_n}(E_x - \mu)^2\right)
\end{aligned}
$$

Hence,

$$g(x,a) = \left(\frac{1}{\sigma_a^2} + \frac{n}{\sigma^2}\right)(a - \hat{a})^2 + g(x,\hat{a}) \tag{9}$$

$$g(x,\hat{a}) = \frac{n}{\sigma^2}\left(V_x + \frac{\sigma^2}{\sigma^2 + n\sigma_a^2}(E_x - \mu)^2\right) \tag{10}$$

The conditional distribution of the shift a given the segment data x, is therefore Gaussian with mean and variance values given by,

$$E(a|x) = \hat{a} \quad , \quad Var(a|x) = V_n^{-1} \tag{11}$$

Substituting (8) and (9), (10) in (2) and carrying out the integration operation results in

$$f(x) = \left(\frac{\sigma^2}{\sigma^2 + n\sigma_a^2}\right)^{\frac{1}{2}} \left(\frac{1}{\sqrt{2\pi}\sigma}\right)^n e^{-\frac{1}{2}g(x,\hat{a})} \tag{12}$$

Comparison of (8) and (12) reveals the following relationship

$$f(x) = \left(\frac{1}{\sigma_a^2} + \frac{n}{\sigma^2}\right)^{-\frac{1}{2}} \sqrt{2\pi}\, f(x,\hat{a}) \tag{13}$$

As can be seen from (13), the likelihood score provided by the approximated MAP method is identical to the true likelihood, except for a term which depends on the length of the segment and not on the segment data.

When $n\sigma_a^2 \ll \sigma^2$, we have $\sigma_a^2 V_n \approx 1$, so that $f(x)$ is reduced to

$$f(x) \approx \left(\frac{1}{\sqrt{2\pi}\sigma}\right)^n \exp\left\{-\frac{1}{2\sigma^2}\sum_t (x_t - \mu)^2\right\}$$

Hence, in that case, the RTSM degenerates to the deterministic model $x_t \sim N(\mu, \sigma^2)$. In the deterministic model, the PDF assigns equal weight to the empirical variance of the samples and

to the distance of the samples empirical average from the grand mean. On the other hand, from (10) it can be seen that the RTSM assigns larger weight to the empirical variance.

We now discuss the problem of parameter estimation for the static RTSM. We discuss two general estimation schemes. The first is a generalized Baum-Welch scheme which is a special case of the expectation - maximization (EM) algorithm (Dempster, Laird and Rubin, 1977). Each iteration is composed of two steps. In the first step (E-step), the conditional expectation of the likelihood of some 'complete data', given the observed data is evaluated. In the second step (M-step), an updated parameters set is obtained, such that the conditional expectation at the new set attains its maximal value. The resulting Baum-Welch algorithm consists of an objective function, in which each state sequence realization, that is consistent with the given measurements, is weighted by its probability of occurrence. At each iteration, the algorithm attempts to bring the objective function to a maximum. The other estimation scheme is an approximation of the Baum-Welch method, based on an extension of the segmental K-means algorithm (Juang and Rabiner, 1990). Each iteration is composed of two steps. In the first stage, a Viterbi decoding algorithm is applied to obtain the most likely state sequence (using the current values of the estimated parameters). Once the segments boundaries are determined, in the second step of the iterative algorithm, the segmental model parameters are re-estimated. Following Ostendorf, Digalakis and Kimball (1996), we have chosen to adopt segmental K-means notation in order to simplify the presentation. We note however, that all the algorithms that will be presented in the sequel can be plugged into the re-estimation step of the Baum-Welch scheme. Note also that the various segmental K-means algorithms that will be presented, iteratively derive optimal unit segmentation as a by-product of the algorithm.

Since the Viterbi stage of the segmental K-means algorithm is standard, we only discuss the re-estimation stage (the second phase of each iteration). Let $x = \left(x^1, ..., x^k\right)$ be k segment realizations associated with the state s. Denote the length of the sequence x^i by n_i. The frames of the segment x^i are denoted by $x_0^i, \ldots, x_{n_i-1}^i$. We wish to obtain the maximum likelihood (ML)

estimates of the model parameters $\mu, \sigma_a^2, \sigma^2$. Setting the derivative of $\log f(x)$ in (12), with respect to μ to zero, we have,

$$\frac{\partial \log f(x)}{\partial \mu} = \sum_{i=1}^{k} -\frac{n_i}{2\sigma^2\sigma_a^2 V_{n_i}} \frac{\partial}{\partial \mu}(E_{x^i} - \mu)^2 = \sum_{i=1}^{k} \frac{n_i}{\sigma^2 + n_i\sigma_a^2}(E_{x^i} - \mu) = 0$$

Hence,

$$\hat{\mu} = \frac{\sum_{i=1}^{k} \frac{1}{\sigma^2 + n_i\sigma_a^2} \sum_{t=0}^{n_i-1} x_t^i}{\sum_{i=1}^{k} \frac{n_i}{\sigma^2 + n_i\sigma_a^2}}$$

When all segments have the same length, denoted by n, the expression above is reduced to

$$\hat{\mu} = \frac{1}{nk} \sum_{i=1}^{k} \sum_{t=0}^{n-1} x_t^i \tag{14}$$

In this case we can also obtain the following closed form expressions for $\hat{\sigma}^2$ and $\hat{\sigma}_a^2$:

$$\hat{\sigma}^2 = \frac{n}{(n-1)k} \sum_{i=1}^{k} V_{x^i} \tag{15}$$

$$\hat{\sigma}_a^2 = \frac{1}{k} \sum_{i=1}^{k} (E_{x^i} - \hat{\mu})^2 - \frac{\hat{\sigma}^2}{n} \tag{16}$$

However, in the general case, where the segment realizations differ in length, we cannot obtain a closed form expression for the ML estimators of σ_a^2 and σ^2, unless some approximation, such as $n_i\sigma_a^2 \gg \sigma^2$ is used (Gales and Young, 1993).

Russell (1993) used the joint probability of the observations and the optimal trajectory as the target function for the maximization problem. Setting the first partial derivatives of $f(x, \hat{a})$ with respect to the estimated parameters to zero yields,

$$\hat{\mu} = \frac{\sum_i \frac{1}{\sigma^2 + n_i\sigma_a^2} \sum_t x_t^i}{\sum_i \frac{n_i}{\sigma^2 + n_i\sigma_a^2}}$$

$$\hat{\sigma}_a^2 = \frac{1}{k} \sum_i (\hat{a}_i)^2$$

$$\hat{\sigma}^2 = \frac{1}{\sum_i n_i} \sum_{i,t} (x_t^i - \hat{a}_i - \hat{\mu})^2$$

Note that $\hat{a}$, which was obtained in (5), is a function of the unknown parameters. Hence, the un-

known parameters appear on both sides of each equation. Russell suggested an iterative solution, where the parameters estimated in the previous iteration of the segmental K-means algorithm are substituted back in the right-hand sides of the equations.

Digalakis, Rohlicek and Ostendorf (1993) considered the static RTSM (which they referred to as a 'target state' segment model) as a special case of the dynamical system model. They suggested to utilize the EM algorithm as an iterative procedure for solving the maximization problem required in the second stage of the segmental K-means algorithm. The hidden values of the random shifts, which are sampled for each segment realization, are the missing data of the EM. A detailed derivation of the EM algorithm for a more general RTSM can be found in Appendix B. Note that if a Baum-Welch algorithm is used, instead of segmental K-means, then the resulting algorithm would involve two levels of EM. In that case, the algorithm in Appendix B, would be in the inner level of the iterations.

3 Scaled Random Trajectory Segment Models

Holmes and Russell (1996) have pointed out that there is a balancing problem between the extra- and the intra-segmental components of the RTSM. Different explanations of an utterance using different number of segments will use different number of extra-segmental probabilities. Therefore, interpretations of the data which involve a large number of short segments require more probability terms than ones which use a small number of long segments. This phenomenon, which does not exist in the standard HMM formalism, is caused by the random segmental element of the RTSM. Holmes and Russell (1996) observed that including self loop transitions in the segment model improves recognition performance. Self loops allow freedom in representing each occurrence of a basic phonetic unit using an optimal number of segments. In this manner the two model components can be automatically balanced. Their preferred solution is, however, to model the intra-segmental variability more accurately. They have found that using a Richter instead of a

Gaussian distribution can greatly improve the performance. In this section we present yet another solution to this balancing problem.

The RTSM can be analyzed from another point of view. As was noted in the previous section, unlike standard HMM, RTSMs assign different weights in the PDF to the empirical variance of the samples and to the distance of the empirical mean of the samples from the grand mean. Recalling (12) and (10), this weighting ratio is given by

$$1 + \frac{\sigma_a^2}{\sigma^2} n$$

This ratio reflects the relative contribution of the extra- and the intra-segmental components in the likelihood function. Note that this ratio depends on the segment duration, n. Therefore, balancing problems can exist even in cases of interpretations of the data which involve the same number of segments but with different segment lengths.

We now suggest a modification of the static RTSM which aims to solve this balancing problem. In this model the ratio between the empirical variance of the samples and the distance of the empirical mean of the samples from the grand mean is independent of the segment duration. This model, which we have termed scaled RTSM, is similar to Russell's model (Russell, 1993) that was presented in the previous section, except that

$$a \sim N\left(0, \frac{\sigma_a^2}{n}\right)$$

where n is the segment length. The scaled RTSM asserts that the variance of the random mean trajectory is inversely proportional to the segment length. This assumption can be justified by the reasoning that short segments are likely to be more affected by their contextual environment than are long ones, due to the coarticulation effect. To assess this assumption, triphone realizations, that were extracted from the *Wall Street Journal* data base, using the SRI DECIPHER system (Digalakis, Monaco and Murveit, 1996), were considered. The various segment realizations were

clustered into groups based on duration, such that all elements within a group had the same duration. As was mentioned in the previous section, in cases where all the segments have the same length, there is no need for using the methods we have described. Instead, the closed form ML solution (14) - (16) can be used. Hence, we were able to estimate the variance of the random trajectory σ_a^2, and the variance of the samples given the trajectory σ^2, separately for each group. Therefore, there was no pre-imposed assumption on the dependence of the parameters on the segment length. This experiment enables us to obtain an empirical dependence of σ_a^2 on the duration, n. We note that the number of segment realizations for each segment was large enough to obtain a reliable estimate to σ_a^2.

Figure 1 shows the reciprocal of the variance of the sampled trajectory σ_a^2, as a function of segment duration, for the first seven mel-cepstrum features, based on 750 realizations of the phoneme 't' in the triphone context 's-t-ih'. It can be seen from Figure 1 that the variance of the sampled trajectory σ_a^2 is inversely proportional to the segment length.

(Figure 1 should be placed about here)

The PDF of an observation segment x in the scaled model is obtained by substituting $\frac{\sigma_a^2}{n}$ in place of σ_a^2 in (12), thus yielding

$$f(x) = \left(\frac{\sigma^2}{\sigma_a^2+\sigma^2}\right)^{\frac{1}{2}} \left(\frac{1}{\sqrt{2\pi}\sigma}\right)^n \exp\left\{-\frac{n}{2\sigma^2}\left(V_x + \frac{\sigma^2}{\sigma_a^2+\sigma^2}(E_x-\mu)^2\right)\right\} \tag{17}$$

We now discuss the parameter estimation problem. A distinct advantage of the scaled model is that we can solve the likelihood equations analytically and do not need to use an iterative algorithm (e.g. EM) for that purpose. Let $x = (x^1, ..., x^k)$ be k segment realizations associated with the state s. Denote the length of the sequence x^i by n_i. To obtain the ML estimators for the model parameters, we set the partial derivatives of $f(x)$, (17) to zero as follows.

$$\frac{\partial \log f(x)}{\partial \mu} = \frac{1}{\sigma_a^2+\sigma^2}\sum_i n_i(E_{x^i}-\mu) = 0$$

$$\frac{\partial \log f(x)}{\partial \sigma_a^2} = -\frac{1}{2(\sigma_a^2+\sigma^2)} \sum_i \left(1 - \frac{1}{\sigma_a^2+\sigma^2} n_i (E_{x^i} - \mu)^2\right) = 0$$

Hence,

$$\sigma_a^2 + \sigma^2 = \frac{1}{k} \sum_i n_i (E_{x^i} - \mu)^2 \tag{18}$$

$$\begin{aligned} \frac{\partial \log f(x)}{\partial \sigma^2} &= -\frac{1}{2\sigma^2} \sum_i (n_i - 1) + \frac{1}{2\sigma^4} \sum_i n_i V_{x^i} - \\ & \frac{1}{2(\sigma_a^2+\sigma^2)} \left(k - \frac{1}{\sigma_a^2+\sigma^2} \sum_i n_i (E_{x^i} - \mu)^2\right) \end{aligned} \tag{19}$$

Substituting (18) in (19) yields closed form solutions to the likelihood equations :

$$\begin{aligned} \hat{\mu} &= \frac{\sum_{i,t} x_t^i}{\sum_i n_i} \\ \hat{\sigma}^2 &= \frac{\sum_i n_i V_{x^i}}{\sum_i (n_i - 1)} \\ \hat{\sigma}_a^2 &= \frac{1}{k} \sum_i n_i (E_{x^i} - \hat{\mu})^2 \quad - \quad \hat{\sigma}^2 \end{aligned}$$

For purpose of comparison we rewrite the re-estimation formulas of the mean in the scaled and non-scaled RTSMs :

$$\begin{aligned} \text{non-scaled model :} \quad \hat{\mu} &= \frac{\sum_i \frac{1}{\sigma^2 + n_i \sigma_a^2} \sum_t x_t^i}{\sum_i \frac{n_i}{\sigma^2 + n_i \sigma_a^2}} \\ \text{scaled model :} \quad \hat{\mu} &= \frac{\sum_{i,t} x_t^i}{\sum n_i} \end{aligned} \tag{20}$$

As can be seen, the re-estimation equation of the non-scaled model assigns smaller weight to frames that correspond to segments with longer duration. On the other hand, the scaled RTSM assigns equal weight to each frame, independently of the duration of the segment that corresponds to that frame. Hence, the re-estimation equation of the scaled model coincides with our intuition that each data sample encapsulates the same amount of information about the mean trajectory.

The scaled model also possesses a computational advantage over the non-scaled model. In order to compute the likelihood of a given utterance, the log-PDF values of the segments in that

utterance need to be summed up. Now, (12) shows that in the non-scaled model, it is required to compute the logarithm of the term $\sigma^2/(\sigma^2 + n\sigma_a^2)$, which depends on n. On the other hand, in the scaled model, the corresponding term, $\sigma^2/(\sigma^2 + \sigma_a^2)$, is independent on n, and may therefore be computed in advance. By assuming a plausible range of segment durations, the duration-dependent variance terms obtained in the unscaled model can be still computed in advance, but there are more terms to compute and therefore also more to store.

4 Scaled Linear Random Trajectory Segment Models

The assumption that the mean trajectory within a state is constant over time, is shared both by the Gaussian HMM and by the static RTSM. In practice, most states violate this assumption. A simple parametric extension of static models is obtained by representing the mean trajectory as a linear function of time.

Deng et al. (1994) proposed a segment model which generalized the standard Gaussian HMM. In their model the mean trajectory is a deterministic linear function of time. In this linear HMM an observation sequence within a state is generated according to :

$$x_t = \mu_a + \mu_b(\frac{t}{n-1} - \frac{1}{2}) + \epsilon_t \qquad t = 0, ..., n-1$$

such that the time index t is initialized to zero at the beginning of the state and then incremented with each new incoming data frame. The linear trajectory is represented here via the line mid point μ_a and the slope μ_b which are state dependent parameters.

Deng and Aksmanovic (1997) extended this linear model by allowing a discrete mixture of linear functions. Holmes and Russell (1995) presented a continuous stochastic variant of a linear HMM. In their model, the linear mean trajectory is a random variable which is sampled on each arrival at the state. The long term variation in this model is represented by an ensemble of linear mean trajectories. The short term variation is considered to be a result of random fluctuation as

it is in the static case. In this model the segmented data is generated according to :

$$x_t = \mu_a + a + (\mu_b + b)(\frac{t}{n-1} - \frac{1}{2}) + \epsilon_t \qquad t = 0, ..., n-1$$

where $x_0, ..., x_{n-1}$ is the observation sequence, μ_a and μ_b are fixed parameters , a and b are independent normal random variables :

$$a \sim N(0, \sigma_a^2) \quad , \quad b \sim N(0, \sigma_b^2)$$

and ϵ_t is a Gaussian white noise term, $\epsilon_t \sim N(0, \sigma^2)$. The line $\mu_a + \mu_b(\frac{t}{n-1} - \frac{1}{2})$ is the average trajectory over all segment realizations and σ_a and σ_b define a distribution function over all linear trajectories. Denote,

$$F_a(n) = n \quad , \quad F_b(n) = \frac{n(n+1)}{12(n-1)} = \sum_{t=0}^{n-1}(\frac{t}{n-1} - \frac{1}{2}) \tag{21}$$

$$E_{a,x} = \frac{1}{F_a(n)} \sum_t x_t \quad , \quad E_{b,x} = \frac{1}{F_b(n)} \sum_t x_t(\frac{t}{n-1} - \frac{1}{2}) \tag{22}$$

$$(\hat{a}, \hat{b}) = \arg\max_{a,b} f(x, a, b)$$

In Appendix A we compute explicit expressions for $\hat{a}$ and $\hat{b}$ and prove that the true PDF $f(x)$ and the MAP approximation $f(x, \hat{a}, \hat{b})$ are related via

$$f(x) = \left(\frac{1}{\sigma_a^2} + \frac{F_a(n)}{\sigma^2}\right)^{-\frac{1}{2}} \left(\frac{1}{\sigma_b^2} + \frac{F_b(n)}{\sigma^2}\right)^{-\frac{1}{2}} 2\pi f(x, \hat{a}, \hat{b})$$

This relation corresponds to (13) in the static case.

As in the static case, there is no closed form expression for the ML estimation of the linear model parameters. Holmes and Russell (1995) proposed an approximated solution which is an extension of their solution for the static RTSM. Alternatively, although the linear RTSM is out of the scope of dynamical system models, the EM approach for static RTSM, suggested by Digalakis (1992), can be generalized to the linear case. The missing data in the EM algorithm are the hidden values of the random variables a and b, which are sampled for each segment realization.

The proposed EM algorithm is developed in Appendix B. The balancing problem, discussed in the previous section, also exists in the linear model. The approach of using a Richter distribution for better modeling the intra-segmental variability, was applied to the linear RTSM by Holmes and Russell (1997).

We now present the scaled version of the linear RTSM. The motivation for this model is similar to that for the static case. The scaled model spreads the information on the hidden linear trajectory uniformly over time. The segment x is generated in the scaled model according to :

$$x_t = \mu_a + a + (\mu_b + b)(\frac{t}{n-1} - \frac{1}{2}) + \epsilon_t \qquad t = 0, ..., n-1$$

The difference from the non-scaled linear model is that now the variances of a and b are dependent on segment duration as follows :

$$a \sim N\left(0, \frac{\sigma_a^2}{F_a(n)}\right) \quad , \quad b \sim N\left(0, \frac{\sigma_b^2}{F_b(n)}\right)$$

The joint PDF of x, a and b is :

$$f(x,a,b) = \frac{\sqrt{F_a(n)}}{\sqrt{2\pi}\sigma_a}\frac{\sqrt{F_b(n)}}{\sqrt{2\pi}\sigma_b}\left(\frac{1}{\sqrt{2\pi}\sigma}\right)^n e^{-\frac{1}{2}g(x,a,b)}$$

where

$$g(x,a,b) = \frac{F_a(n)a^2}{\sigma_a^2} + \frac{F_b(n)b^2}{\sigma_b^2} + \frac{1}{\sigma^2}\sum_t (x_t - (\mu_a + a) - (\mu_b + b)(\frac{t}{n-1} - \frac{1}{2}))^2$$

Algebraic manipulation of $g(x,a,b)$ reveals:

$$\begin{aligned}(\hat{a}, \hat{b}) &= \arg\max_{a,b} f(x,a,b) = E((a,b)|x) \\ &= \left(\frac{\frac{1}{\sigma^2}\sum_t (x_t - \mu_a)}{F_a(n)(\frac{1}{\sigma_a^2} + \frac{1}{\sigma^2})} \quad , \quad \frac{\frac{1}{\sigma^2}\sum_t (x_t - \mu_b(\frac{t}{n-1} - \frac{1}{2}))(\frac{t}{n-1} - \frac{1}{2})}{F_b(n)(\frac{1}{\sigma_b^2} + \frac{1}{\sigma^2})}\right)\end{aligned} \tag{23}$$

To gain further insight to the probabilistic behavior of the model we derive an explicit expression for $f(x)$. In Appendix A we have computed the following equivalent expression for $g(x,a,b)$

$$g(x,a,b) = \left(\frac{F_a(n)}{\sigma_a^2} + \frac{F_a(n)}{\sigma^2}\right)(a - \hat{a})^2 + \left(\frac{F_b(n)}{\sigma_b^2} + \frac{F_b(n)}{\sigma^2}\right)(b - \hat{b})^2 + g(x, \hat{a}, \hat{b}) \tag{24}$$

and $g(x,\hat{a},\hat{b})$ may be written as:

$$g(x,\hat{a},\hat{b}) = \frac{1}{\sigma^2}(\sum_t x_t^2 \quad + \quad F_a(n)(\frac{\sigma^2}{\sigma_a^2+\sigma^2}(E_{a,x}-\mu_a)^2 - (E_{a,x})^2) \tag{25}$$
$$+ \quad F_b(n)(\frac{\sigma^2}{\sigma_b^2+\sigma^2}(E_{b,x}-\mu_b)^2 - (E_{b,x})^2))$$

The following explicit expression for $f(x)$ results,

$$f(x) = \left(\frac{\sigma^2}{\sigma_a^2+\sigma^2}\right)^{\frac{1}{2}} \left(\frac{\sigma^2}{\sigma_b^2+\sigma^2}\right)^{\frac{1}{2}} \left(\frac{1}{\sqrt{2\pi}\sigma}\right)^n e^{-\frac{1}{2}g(x,\hat{a},\hat{b})} \tag{26}$$

We now discuss the parameter estimation problem given the segmented data. In spite of the fact that the linear model is more complicated than the static one, closed form expressions for the estimated parameters may be derived. Let $x^1,...,x^k$ be k segment realizations associated with the state s. Denote the length of the sequence x^i by n_i. The ML equations are :

$$\frac{\partial \log f(x)}{\partial \sigma_a^2} = \frac{1}{\sigma_a^2+\sigma^2}\sum_i \left(1 - \frac{1}{\sigma_a^2+\sigma^2}F_a(n_i)(E_{a,x^i}-\mu_a)^2\right) = 0$$

Hence,

$$\sigma_a^2+\sigma^2 = \frac{1}{k}\sum_i F_a(n_i)(E_{a,x^i}-\mu_a)^2 \tag{27}$$

$$\frac{\partial \log f(x)}{\partial \sigma_b^2} = \frac{1}{\sigma_b^2+\sigma^2}\sum_i \left(1 - \frac{1}{\sigma_b^2+\sigma^2}F_b(n_i)(E_{b,x^i}-\mu_b)^2\right) = 0$$

Hence,

$$\sigma_b^2+\sigma^2 = \frac{1}{k}\sum_i F_b(n_i)(E_{b,x^i}-\mu_b)^2 \tag{28}$$

$$\begin{aligned}\frac{\partial \log f(x)}{\partial \sigma^2} \quad = \quad & \frac{1}{\sigma^2}\sum_i (n_i-2) \\ & -\frac{1}{\sigma^4}\sum_i(\sum_t (x_t^i)^2 - F_a(n_i)(E_{a,x^i})^2 - F_b(n_i)(E_{b,x^i})^2) \\ & +\frac{1}{\sigma_a^2+\sigma^2}\left(k - \frac{1}{\sigma_a^2+\sigma^2}\sum_i F_a(n_i)(E_{a,x^i}-\mu_a)^2\right) \\ & +\frac{1}{\sigma_b^2+\sigma^2}\left(k - \frac{1}{\sigma_b^2+\sigma^2}\sum_i F_b(n_i)(E_{b,x^i}-\mu_b)^2\right)\end{aligned} \tag{29}$$

Substituting (27) and (28) in (29) enables us to solve the likelihood equations. Therefore the ML estimators are :

$$
\begin{aligned}
\hat{\mu}_a &= \frac{\sum_i F_a(n_i) E_{a,x^i}}{\sum_i F_a(n_i)} = \frac{\sum_{i,t} x_t^i}{\sum_i F_a(n_i)} \\
\hat{\mu}_b &= \frac{\sum_i F_b(n_i) E_{b,x^i}}{\sum_i F_b(n_i)} = \frac{\sum_{i,t} x_t^i (\frac{t}{n_i - 1} - \frac{1}{2})}{\sum_i F_b(n_i)} \\
\hat{\sigma}^2 &= \frac{\sum_i (\sum_t (x_t^i)^2 - F_a(n_i)(E_{a,x^i})^2 - F_b(n_i)(E_{b,x^i})^2)}{\sum_i (n_i - 2)} \\
\hat{\sigma}_a^2 &= \frac{1}{k} \sum_i F_a(n_i)(E_{a,x^i} - \hat{\mu}_a)^2 \quad - \quad \hat{\sigma}^2 \\
\hat{\sigma}_b^2 &= \frac{1}{k} \sum_i F_b(n_i)(E_{b,x^i} - \hat{\mu}_b)^2 \quad - \quad \hat{\sigma}^2
\end{aligned}
$$

5 Experimental Results

We evaluated the model presented in the previous section using the ARPA, large vocabulary, speaker independent, continuous speech, *Wall Street Journal* (WSJ) corpus. Experiments were conducted with DECIPHER, SRI's continuous speech recognition system (Digalakis, Monaco and Murveit, 1996). The recognizer was configured with a front-end that outputs a 39-dimensional vector. The first components of the vector consist of 12 cepstral coefficients and an energy term. The other components of the feature vector are the first and second time derivatives of the first 13 components.

The task we choose for evaluation is phonetic classification. In classification the correct segmentation (phone beginning and ending time) of the input observation sequence is given. Our objective is to assign correct phone labels to each segment. The DECIPHER system was used to determine automatically the phone segmentation for each sentence in the database. Having obtained phonetically aligned test data, the actual classification process is just a matter of finding the most likely phone label for a speech segment according to the models being evaluated. The training set consists of 100 realizations in various contexts for each phone. The testing set consists of another 100 realizations for each phoneme.

The goal of the experiments we have conducted is to compare between the performance of the scaled and unscaled random mean trajectory models. In previous sections we have discussed both the cases of constant and linear mean trajectories. The experiments were performed for static as well as linear models. This enable us to find how much the assumption of linear mean trajectory improves the performance. We also add results for models where the mean trajectory is a deterministic parameter (i.e. standard Gaussian HMM and Deng's linear model) as a reference. It should be noted that a deterministic model has less parameters than the corresponding random model. The acoustic models we implemented for evaluation were:

1. Standard Gaussian HMM.
2. Static RTSM (Russell, 1993).
3. Scaled static RTSM (presented in section 3).
4. Linear mean trajectory segment model (Deng et al., 1994).
5. Linear RTSM (Holmes and Russell, 1995).
6. Scaled linear RTSM (presented in section 4).

Several alternatives for model topologies were employed in order to analyze the balancing problem in different situations. The first analyzed topology assigns one segment to the entire phone. The second one still associates a single segment model with each phone, but includes self loops. In other words, the phone can be modeled by a number of segments which are all corresponding to the same segmental model. The third topology, models each phone using three states. The last examined topology is also a three state model but a skip over a state is allowed. In this manner a phone can be explained using at most three states. Note that in the second and the forth topologies the number of segments per phone is not fixed. Balancing the model components in those cases is critical.

Training the models that include more than one segment per phone was done using the segmental K-means algorithm. The parameters of the unscaled random models were computed using the iterative inner EM algorithm that is presented in appendix B. We preferred to use this training method because, although it is an iterative procedure, it computes the true likelihood function. The scaled model was trained using the closed from formulae that were developed in previous sections. The random models, both scaled and unscaled, were initialized using the deterministic version of the model for the first iteration of the segmental K-means algorithm. Duration was not modeled explicitly. Therefore, all durations were assigned equal probability.

(Table 1 should be placed about here)

As can be seen, the scaled model usually outperforms the previously suggested non-scaled model, both for the static case and for the linear case. These results also reassess the significant performance improvement caused by using a linear model instead of a static one. It is noteworthy to mention that by comparing the parameters of the scaled and the corresponding unscaled model, we have found that the values of the mean parameters and the values of the intra-segmental variances are similar.

Another task we evaluated is triphone classification. Given a triphone context (the phones before and after the current one), the goal is to determine the label of the current phone based on the acoustics. Context dependent classification was chosen because, in that case there are fewer discrepancies between utterances. Hence, in practice, this is usually the case of interest when using segment models. The topology we have used for this task is the simplest one. Each phoneme is modeled by a single segmental model. This is the first topology from the topologies list of the previous experiment.

In Table 2 we present recognition results for some frequently occurring triphone contexts. The first column in Table 2 summarizes a classification experiment given the triphone context m-n. The segmented WSJ database was used to extract the phones that appear between the phones m and n. The classification was done among those phonemes that have a significant number of

occurrences (at least 60) in the m-n context. In this context these phonemes are (using ARPABET notation) : aa, ae, ah, aw, ax, ay, eh, ey and iy. Half of the data was used to train the triphone models. The other half was used for the actual classification task. The subsequent four columns present results for similar classification tasks in other triphone contexts. The final column presents the classification performance averaged over the 120 most frequently occurring triphone contexts. As can be seen, the scaled model outperforms the previously suggested non scaled model.

(Table 2 should be placed about here)

6 Conclusion

In this study we have proposed, implemented and evaluated a new type of random trajectory segment model where the variance of the mean trajectory is inversely proportional to the segment duration. In this model the division of the acoustic information in an utterance does not depend on a specific segmentation. Instead, we extract the same amount of information about the mean trajectory from each data frame. We have named this approach a scaled RTSM. One desirable attribute of the scaled model is that it leads to a simple training algorithm. More precisely, given a training set, consisting of a list of segment realizations, the ML estimation of the scaled model can be solved analytically. On the other hand, in the non-scaled model, an iterative algorithm, (e.g., EM) is required. Such an iterative algorithm, is not guaranteed to reach the global maximum.

Appendix A

We derive an explicit expression for the PDF of the linear RTSM. Both the scaled and the non-scaled versions are discussed. We begin by analyzing the scaled model. The definition of the scaled linear model implies that the joint PDF of the observed segment x of length n and the linear function coefficients a and b is :

$$f(x,a,b) = \frac{\sqrt{F_a(n)}}{\sqrt{2\pi}\sigma_a}\frac{\sqrt{F_b(n)}}{\sqrt{2\pi}\sigma_b}\left(\frac{1}{\sqrt{2\pi}\sigma}\right)^n e^{-\frac{1}{2}g(x,a,b)}$$

where

$$g(x,a,b) = \frac{F_a(n)a^2}{\sigma_a^2} + \frac{F_b(n)b^2}{\sigma_b^2} + \frac{1}{\sigma^2}\sum_t (x_t - (\mu_a + a) - (\mu_b + b)(\frac{t}{n-1} - \frac{1}{2}))^2$$

We show that $g(x,a,b)$ may be written as:

$$\begin{aligned} g(x,a,b) &= \left(\frac{F_a(n)}{\sigma_a^2} + \frac{F_a(n)}{\sigma^2}\right)(a-\hat{a})^2 + \left(\frac{F_b(n)}{\sigma_b^2} + \frac{F_b(n)}{\sigma^2}\right)(b-\hat{b})^2 + \\ &\quad \frac{1}{\sigma^2}(F_a(n)(\frac{\sigma^2}{\sigma_a^2+\sigma^2}(E_{a,x}-\mu_a)^2 - (E_{a,x})^2) + \\ &\quad\quad F_b(n)(\frac{\sigma^2}{\sigma_b^2+\sigma^2}(E_{b,x}-\mu_b)^2 - (E_{b,x})^2) + \sum_t x_t^2) \end{aligned}$$

The terms $\hat{a}$, $\hat{b}$, $F_a(n)$, $F_b(n)$, $E_{a,x}$ and $E_{b,x}$ are defined in (21), (22).

$$\begin{aligned} g(x,a,b) &= \frac{F_a(n)a^2}{\sigma_a^2} + \frac{F_b(n)b^2}{\sigma_b^2} + \frac{1}{\sigma^2}\sum_t (x_t - (\mu_a + a) - (\mu_b + b)(\frac{t}{n-1} - \frac{1}{2}))^2 \\ &= \left(\frac{F_a(n)}{\sigma_a^2} + \frac{F_a(n)}{\sigma^2}\right)a^2 - 2a\frac{1}{\sigma^2}\sum_t (x_t - \mu_a) + \\ &\quad \left(\frac{F_b(n)}{\sigma_b^2} + \frac{F_b(n)}{\sigma^2}\right)b^2 - 2b\frac{1}{\sigma^2}\sum_t (x_t - \mu_b(\frac{t}{n-1} - \frac{1}{2}))(\frac{t}{n-1} - \frac{1}{2}) + \\ &\quad\quad \frac{1}{\sigma^2}\sum_t (x_t - \mu_a - \mu_b(\frac{t}{n-1} - \frac{1}{2}))^2 \\ &= \left(\frac{F_a(n)}{\sigma_a^2} + \frac{F_a(n)}{\sigma^2}\right)\left(a - \frac{\sigma_a^2}{\sigma_a^2+\sigma^2}(E_{a,x}-\mu_a)\right)^2 + \\ &\quad \left(\frac{F_b(n)}{\sigma_b^2} + \frac{F_b(n)}{\sigma^2}\right)\left(b - \frac{\sigma_b^2}{\sigma_b^2+\sigma^2}(E_{b,x}-\mu_b)\right)^2 \\ &\quad -\frac{F_a(n)}{\sigma^2}\frac{\sigma_a^2}{\sigma_a^2+\sigma^2}(E_{a,x}-\mu_a)^2 - \frac{F_b(n)}{\sigma^2}\frac{\sigma_b^2}{\sigma_b^2+\sigma^2}(E_{b,x}-\mu_b)^2 + \\ &\quad\quad \frac{1}{\sigma^2}\sum_t (x_t - \mu_a - \mu_b(\frac{t}{n-1} - \frac{1}{2}))^2 \end{aligned} \tag{30}$$

Given the observation sequence x, $g(x,a,b)$ is a quadratic form in a and b. Hence, we conclude that the conditional distribution of a and b given the segment x is Gaussian. Furthermore, a and b are conditionally independent. Now, the first moments may be read directly from (30):

$$
\begin{aligned}
E(a|x) &= \hat{a} = \left(\frac{\sigma_a^2}{\sigma_a^2+\sigma^2}\right)(E_{a,x}-\mu_a) \\
Var(a|x) &= \left(\frac{F_a(n)}{\sigma_a^2}+\frac{F_a(n)}{\sigma^2}\right)^{-1} \\
E(b|x) &= \hat{b} = \left(\frac{\sigma_b^2}{\sigma_b^2+\sigma^2}\right)(E_{b,x}-\mu_b) \\
Var(b|x) &= \left(\frac{F_b(n)}{\sigma_b^2}+\frac{F_b(n)}{\sigma^2}\right)^{-1}
\end{aligned}
$$

Direct algebraic manipulations reveal the following relation

$$\sum_t (x_t - \mu_a - \mu_b(\frac{t}{n-1}-\frac{1}{2}))^2 =$$

$$F_a(n)((E_{a,x}-\mu_a)^2-(E_{a,x})^2)+F_b(n)((E_{b,x}-\mu_b)^2-(E_{b,x})^2)+\sum_t x_t^2$$

Substituting this relation in (30) yields :

$$
\begin{aligned}
g(x,a,b) &= \left(\frac{F_a(n)}{\sigma_a^2}+\frac{F_a(n)}{\sigma^2}\right)(a-\hat{a})^2+\left(\frac{F_b(n)}{\sigma_b^2}+\frac{F_b(n)}{\sigma^2}\right)(b-\hat{b})^2+ \\
&\quad \frac{1}{\sigma^2}(\sum x_t^2 + F_a(n)\left((1-\frac{\sigma_a^2}{\sigma_a^2+\sigma^2})(E_{a,x}-\mu_a)^2-(E_{a,x})^2\right) \\
&\quad +F_b(n)\left((1-\frac{\sigma_b^2}{\sigma_b^2+\sigma^2})(E_{b,x}-\mu_b)^2-(E_{b,x})^2\right))
\end{aligned}
$$

Hence,

$$g(x,a,b)=\left(\frac{F_a(n)}{\sigma_a^2}+\frac{F_a(n)}{\sigma^2}\right)(a-\hat{a})^2+\left(\frac{F_b(n)}{\sigma_b^2}+\frac{F_b(n)}{\sigma^2}\right)(b-\hat{b})^2+g(x,\hat{a},\hat{b})$$

Now, $g(x,\hat{a},\hat{b})$ may be written as :

$$
\begin{aligned}
g(x,\hat{a},\hat{b}) = \frac{1}{\sigma^2}(\sum_t x_t^2 &+ F_a(n)(\frac{\sigma^2}{\sigma_a^2+\sigma^2}(E_{a,x}-\mu_a)^2-(E_{a,x})^2) \\
&+ F_b(n)(\frac{\sigma^2}{\sigma_b^2+\sigma^2}(E_{b,x}-\mu_b)^2-(E_{b,x})^2))
\end{aligned}
$$

Using this representation we may solve the double integral :

$$f(x)=\int_a\int_b f(x,a,b)\,da\,db$$

and obtain the following explicit expression for the PDF,

$$f(x) = \left(\frac{F_a(n)}{\sigma_a^2}+\frac{F_a(n)}{\sigma^2}\right)^{-\frac{1}{2}}\left(\frac{F_b(n)}{\sigma_b^2}+\frac{F_b(n)}{\sigma^2}\right)^{-\frac{1}{2}}2\pi f(x,\hat{a},\hat{b})$$

$$= \left(\frac{\sigma^2}{\sigma_a^2+\sigma^2}\right)^{\frac{1}{2}} \left(\frac{\sigma^2}{\sigma_b^2+\sigma^2}\right)^{\frac{1}{2}} \left(\frac{1}{\sqrt{2\pi}\sigma}\right)^n e^{-\frac{1}{2}g(x,\hat{a},\hat{b})}$$

The expressions we derived for $f(x)$, $g(x,a,b)$ and the moments of the conditional distribution of a and b given x may be easily transformed to the non-scaled linear case. We need only substitute $F_a(n)\sigma_a^2$ and $F_b(n)\sigma_b^2$ for σ_a^2 and σ_b^2. For example, in the non-scaled case we have :

$$\begin{aligned} E(a|x) &= \hat{a} = \left(\frac{F_a(n)\sigma_a^2}{F_a(n)\sigma_a^2+\sigma^2}\right)(E_{a,x}-\mu_a) \\ E(b|x) &= \hat{b} = \left(\frac{F_b(n)\sigma_b^2}{F_b(n)\sigma_b^2+\sigma^2}\right)(E_{b,x}-\mu_b) \end{aligned}$$

and the PDF of the observation sequence x of length n is :

$$f(x) = \left(\frac{1}{\sigma_a^2}+\frac{F_a(n)}{\sigma^2}\right)^{-\frac{1}{2}} \left(\frac{1}{\sigma_b^2}+\frac{F_b(n)}{\sigma^2}\right)^{-\frac{1}{2}} 2\pi f(x,\hat{a},\hat{b})$$

Appendix B

The static RTSM can be seen as a special case of the dynamical system model studied by Digalakis (1992) such that the state space is constant over time. The EM algorithm presented by Digalakis (1992) can be applied to the static RTSM. In this appendix we derive the EM re-estimation equations for the unscaled linear RTSM which is out of the scope of dynamical system models. The re-estimation equations for static RTSM can be easily deduced from the equations developed here. Suppose we have k segment realizations $x^1,...,x^k$. Denote the length of x^i by n_i. Denote by a_i and b_i the hidden coefficients of the line which was sampled for the segment x^i. Define $z_i = (x^i, a_i, b_i)$, $\quad z_1,...,z_k$ are the complete data for this EM framework. Denote the parameter set we want to estimate by $\theta = \{\mu_a, \mu_b, \sigma_a^2, \sigma_b^2, \sigma^2\}$. The current estimate at the beginning of the iteration is denoted by $\theta_0 = \{\mu_{a0}, \mu_{b0}, \sigma_{a0}^2, \sigma_{b0}^2, \sigma_0^2\}$.

$$\log f(z_i,\theta) = \log\sigma_a^2 + \frac{a_i^2}{\sigma_a^2} + \log\sigma_b^2 + \frac{b_i^2}{\sigma_b^2} + n_i\log\sigma^2 +$$

$$\frac{1}{\sigma^2}\sum_t (x_t^i - (\mu_a - a_i) - (\mu_b - b_i)(\frac{t}{n-1} - \frac{1}{2}))^2 + C$$

where C is a constant that is independent of the parameter vector θ. Hence,

$$E(\log f(z_i, \theta)|x, \theta_0) = n_i \log \sigma^2 +$$

$$\log \sigma_a^2 + \frac{1}{\sigma_a^2} E(a_i^2|x^i, \theta_0) + \log \sigma_b^2 + \frac{1}{\sigma_b^2} E(b_i^2|x^i, \theta_0) +$$

$$+ \frac{1}{\sigma^2}\sum_t E((x_t^i - (\mu_a - a_i) - (\mu_b - b_i)(\frac{t}{n_i - 1} - \frac{1}{2}))^2|x^i, \theta_0)$$

Denote :

$$\begin{aligned}
E_{a_i} &= E(a_i|x^i, \theta_0) = \left(\frac{1}{\sigma_{a0}^2} + \frac{F_a(n_i)}{\sigma_0^2}\right)^{-1} \frac{F_a(n_i)}{\sigma_0^2}(E_{a,x^i} - \mu_{a0}) \\
V_{a_i} &= Var(a_i|x^i, \theta_0) = \left(\frac{1}{\sigma_{a0}^2} + \frac{F_a(n_i)}{\sigma_0^2}\right)^{-1} \\
E_{b_i} &= E(b_i|x^i, \theta_0) = \left(\frac{1}{\sigma_{b0}^2} + \frac{F_b(n_i)}{\sigma_0^2}\right)^{-1} \frac{F_b(n_i)}{\sigma_0^2}(E_{b,x^i} - \mu_{b0}) \\
V_{b_i} &= Var(b_i|x^i, \theta_0) = \left(\frac{1}{\sigma_{b0}^2} + \frac{F_b(n_i)}{\sigma_0^2}\right)^{-1}
\end{aligned}$$

where

$$\begin{aligned}
F_a(n_i) &= n_i \quad , \quad E_{a,x^i} = \frac{1}{F_a(n_i)}\sum_t x_t^i \\
F_b(n_i) &= \frac{n_i(n_i+1)}{12(n_i-1)} \quad , \quad E_{b,x^i} = \frac{1}{F_b(n_i)}\sum_t x_t^i(\frac{t}{n_i - 1} - \frac{1}{2})
\end{aligned}$$

These relations are developed in Appendix A. Direct algebraic manipulations reveal the following relation :

$$E((x_t^i - (\mu_a - a_i) - (\mu_b - b_i)(\frac{t}{n_i - 1} - \frac{1}{2}))^2|x^i, \theta_0) =$$

$$F_a(n_i)V_{a_i} + F_b(n_i)V_{b_i} + \sum_t (x_t^i - (\mu_a + E_{a_i}) - (\mu_b + E_{b_i})(\frac{t}{n_i - 1} - \frac{1}{2}))^2$$

We now define the EM auxiliary function :

$$
\begin{aligned}
Q(\theta,\theta_0) &= E(\log f(z,\theta)|x,\theta_0) = \sum_i E(\log f(z_i,\theta)|x^i,\theta_0) \\
&= k\log\sigma_a^2 + \frac{1}{\sigma_a^2}\sum_i E(a_i^2|x^i,\theta_0) + \frac{1}{\sigma^2}\sum_i F_a(n_i)V_{a_i} + \\
&\quad k\log\sigma_b^2 + \frac{1}{\sigma_b^2}\sum_i E(b_i^2|x^i,\theta_0) + \frac{1}{\sigma^2}\sum_i F_b(n_i)V_{b_i} + \\
&\quad \sum_i n_i\log\sigma^2 + \frac{1}{\sigma^2}\sum_{i,t}(x_t^i - (\mu_a + E_{a_i}) - (\mu_b + E_{b_i})(\frac{t}{n_i - 1} - \frac{1}{2}))^2
\end{aligned}
$$

Optimization of the auxiliary function with respect to the model parameters yields the following new estimated parameter vector (this optimization is carried out by setting the corresponding partial derivatives to zero).

$$
\begin{aligned}
\hat{\mu}_a &= \frac{\sum_i F_a(n_i)(E_{a,x^i} - E_{a_i})}{\sum_i F_a(n_i)} \\
\hat{\mu}_b &= \frac{\sum_i F_b(n_i)(E_{b,x^i} - E_{b_i})}{\sum_i F_b(n_i)} \\
\hat{\sigma}_a^2 &= \frac{1}{k}\sum_i (V_{a_i} + E_{a_i}^2) = \frac{1}{k}\sum_i E(a_i^2|x^i,\theta_0) \\
\hat{\sigma}_b^2 &= \frac{1}{k}\sum_i (V_{b_i} + E_{b_i}^2) = \frac{1}{k}\sum_i E(b_i^2|x^i,\theta_0) \\
\hat{\sigma}^2 &= \frac{1}{\sum_i n_i}\sum_i (F_a(n_i)V_{a_i} + F_b(n_i)V_{b_i} + \\
&\qquad \sum_t (x_t^i - (\hat{\mu}_a + E_{a_i}) - (\hat{\mu}_b + E_{b_i})(\frac{t}{n_i - 1} - \frac{1}{2}))^2)
\end{aligned}
$$

References

Dempster, A. P., Laird, N. M. & Rubin, D. B. (1977). Maximum likelihood estimation from incomplete data, *Journal of the Royal Statistics Society*, 39, 1-38.

Deng, L., Aksmanovic, M., Sun, D. & Wu, J. (1994). Speech recognition using hidden Markov models with polynomial regression functions as non stationary states, *IEEE Transactions on Speech and Audio Processing* 2, 507-520.

Deng, L. & Aksmanovic, M. (1997). Speaker independent phonetic classification using hidden

Markov models with state conditioned mixtures of trend functions, *IEEE Transactions on Speech and Audio Processing*, 5, 319-324.

Digalakis, V., Rohlicek, J. & Ostendorf, M. R. (1993). A dynamical system approach to continuous speech recognition. *IEEE Transactions on Speech and Audio Processing*, 1, 431-442.

Digalakis, V. (1992). Segment-based stochastic models of spectral dynamics for continuous speech recognition, Ph.D Thesis, Boston University.

Digalakis, V., Monaco, P. & Murveit, H. (1996). Genones: generalized mixture tying in continuous hidden Markov model-based speech recognizers, *IEEE Transactions on Speech and Audio Processing* 4, 281-289.

Gales, M. & Young, S. (1993). The theory of segmental hidden Markov models, Technical Report CUED/F-INFENG/TR 133, Cambridge, U.K.

Ghitza, O. & Sondhi, M. (1993). Hidden Markov models with templates as non-stationary states: an application to speech recognition, *Computer Speech and Language* 7, 101-119.

Goldberger, J., Burshtein, D. & Franco, H. (1997). Segmental modeling using a continuous mixture of non-parametric models, submitted for publication.

Holmes, W. & Russell, M. (1995). Speech recognition using a linear dynamic segmental HMMs, *Proc. Eurospeech*, pp. 1611-1614.

Holmes, W. & Russell, M. (1996). Modeling speech variability with segmental HMMs, *Proc. Int. Conf. Acoustics, Speech and Signal Processing*, 447-450.

Holmes, W. & Russell, M. (1997). Linear dynamic segmental HMMS : Variability representation and training procedure, *Proc. Int. Conf. Acoustics, Speech and Signal Processing*, 1399-1402.

Juang, B. H. & Rabiner, L. R. (1990). The segmental K-means algorithm for estimating parameters of hidden Markov models, *IEEE Transactions on Acoustics, Speech and Signal Processing* 38, 1639-1641.

Ostendorf, M., Digalakis, V. & Kimball, O. A. (1996). From HMMs to segment models: a

unified view of stochastic modeling for speech recognition, *IEEE Transactions on Speech and Audio Processing* 4, 360-377.

Russell, M. (1993). A segmental HMM for speech pattern modeling, *Proc. Int. Conf. Acoustics, Speech and Signal Processing*, 499-502.

Acknowledgment

We acknowledge the speech group at SRI international for providing us segmented data obtained by using the DECIPHER speech recognition system.

	model	one state	one state self-loop	three states	three states with skips
static	deterministic	52.1		61.7	61.9
	non-scaled	50.7	52.4	61.1	61.7
	scaled	51.5	55.1	62.3	62.8
linear	deterministic	57.2	57.8	61.8	62.3
	non-scaled	58.0	57.0	63.3	63.9
	scaled	59.1	58.1	62.2	64.1

Table 1: Phoneme classification rate results

	model	m-?-n	p-?-t	s-?-ae	t-?-s	k-?-t	average
static	non-scaled	49.4	61.0	80.3	69.3	82.9	79.6
	scaled	52.7	65.1	79.6	71.4	84.2	80.3
linear	non-scaled	59.3	64.0	88.4	72.1	87.5	82.0
	scaled	61.9	67.7	88.9	74.8	90.0	82.4

Table 2: Triphone classification rate results

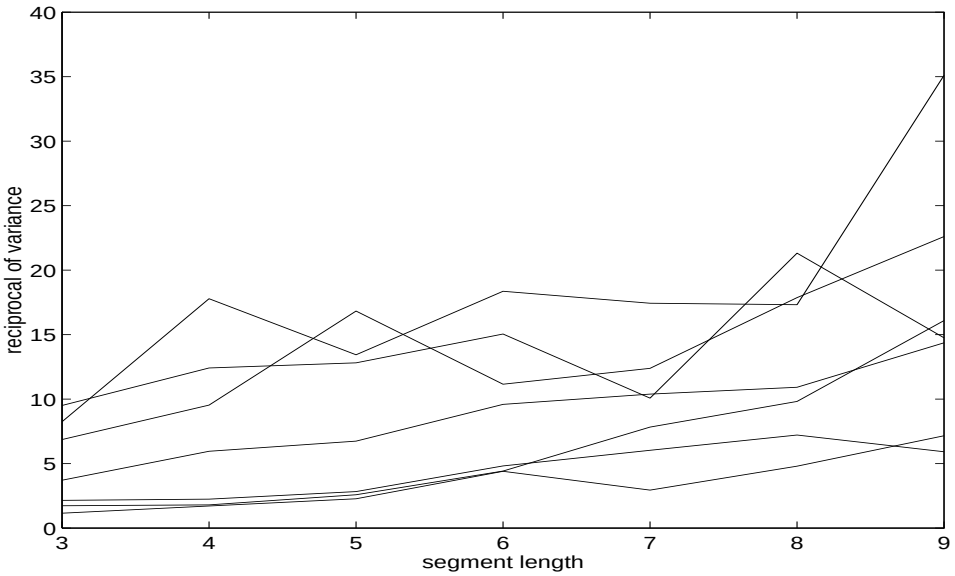

Figure 1: The reciprocal of the variance of the sampled mean trajectory as a function of the segment length, for the first seven mel-cepstrum features, based on 750 realizations of the phoneme 't' in the triphone context 's-t-ih'.