SCALED RANDOM SEGMENTAL MODELS

Jacob Goldberger and David Burshtein

Department of Electrical Engineering - Systems
Tel-Aviv University
Tel-Aviv Israel 69978
jacob,burstyn@eng.tau.ac.il

ABSTRACT

We present the concept of a scaled random segmental model, which aims to overcome the modeling problem created by the fact that segment realizations of the same phonetic unit differ in length. In the scaled model the variance of the random mean trajectory is inversely proportional to the segment length. The scaled model enables a Baum-Welch type parameter reestimation, unlike the previously suggested, non-scaled models, that require more complicated iterative estimation procedures. In experiments we have conducted with phoneme classification, it was found that the scaled model shows improved performance compared to the non-scaled model.

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. These methods are usually known by the name segment models, since the fundamental modeling unit is the entire phonetic unit (segment), unlike HMMs that utilize frame-based modeling. A comprehensive survey on the subject can be found in [6]. Over the past decade a number of studies have proposed segment models composed of stochastic description of the mean trajectory, as an alternative to the multi description of the mean trajectory, that is provided by mixture of Gaussians HMMs. Random trajectory segmental modeling can be thought of a generalization of the Gaussian HMM formalism. The main difference is that the mean trajectory 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 caused by factors 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 trajectory. The short term variability within a state is modeled by the deviation of the feature vectors from the sampled mean trajectory. In standard HMM these two effects are modeled implicitly by a single PDF. In this paper we concentrate on the case where the trajectories PDF is Gaussian and the mean trajectory within the state boundaries is either constant or linear function of time. We term this models static random segmental model and linear random segmental model, respectively. The static model was originally presented by Russell [7] and by Gales and Young [4]. This model is reviewed in the next section. We shall analyze the shortcomings of the static model and suggest a modification of the model through scaling the variance of the trajectories PDF according to the segment length. The linear model was introduced by Holmes and Russell [5]. We shall present the scaled version of the linear model. We show that the scaled models are much easily trained and show better performance.

2. STATIC RANDOM MODEL

A static random segmental model assumes that the observations within an HMM state $x = (x_{t_1}, ..., x_{t_2})$ are realized according to :

$$x_t = \mu_{j,a} + a + \epsilon_t \qquad t = t_1, ..., t_2$$

where $\mu_{j,a}$ is a fixed parameter, associated with the state j, that describes the grand mean trajectory. $t_1, ..., t_2$ is the time interval of the sojourn in the state j. The random variable a is a shift of the mean trajectory. It is sampled at the transition into the state j and is global to the visit in the state. It is assumed that $a \sim N(0, \sigma^2_{j,a})$ (i.e., a is a Gaussian random variable with zero mean and variance $\sigma^2_{j,a}$). The short term variability is represented by ϵ_t, which is a zero mean Gaussian random variable with state dependent variance, $\epsilon_t \sim N(0, \sigma^2_j)$. A closed-form expression for the PDF of the data segment associated with the state j can be obtained as follows [4] :

$$f_j(x, t_1, t_2) = \left(\frac{\sigma^2_j}{\sigma^2_j + \Delta t\, \sigma^2_{j,a}}\right)^{\frac{1}{2}} \left(\frac{1}{2\pi\sigma^2_j}\right)^{\frac{\Delta t}{2}} e^{-\frac{1}{2} g_j(x, t_1, t_2)} \tag{1}$$

$$g_j(x, t_1, t_2) = \frac{\Delta t}{\sigma^2_j} V(x, t_1, t_2) + \frac{\Delta t}{\sigma^2_j + \Delta t\, \sigma^2_{j,a}} (E(x, t_1, t_2) - \mu_{j,a})^2$$

where

$$E(x,t_1,t_2) = \frac{1}{\Delta t}\sum_{t=t_1}^{t_2} x_t$$

$$V(x,t_1,t_2) = \frac{1}{\Delta t}\sum_{t=t_1}^{t_2} x_t^2 - \left(\frac{1}{\Delta t}\sum_{t=t_1}^{t_2} x_t\right)^2$$

and Δt is the segment duration, namely $t_2 - t_1 + 1$. To simplify notation, it is assumed that the observations are one dimensional. Generalization to the multi-dimensional case is straight-forward. A left to right HMM topology is assumed.

According to this model, during the generation of the utterance, first a state sequence is chosen and then the observations sequence is sampled according to that state sequence. The probability of the utterance x is obtained by summing over all possible state sequences :

$$f(x) = \sum_s f(x,s)$$

Dynamic programming must be applied in order to efficiently compute this expression. A similar dynamic programming is performed in order to compute the likelihood score in the standard HMM. However, in segmental models computation of the density function is far more complex. This is due to the fact that the probability of a frame does not depend only on the state but also on the location of this frame within the segment sampled during the visit in the state. In a segmental model we can not compute the probability of a single frame in a state. We must compute the probability of the entire segment. The complexity of the algorithm can be reduced by assuming a maximal state duration.

We now discuss the parameter estimation problem. Assume that the training data-base consists of the k utterances $x^1, ..., x^k$. In the Viterbi decoding approach we estimate the parameters using only the best suited state sequence. For the Baum-Welch algorithm, however, we must consider all the possible state sequences. Each state sequence is considered according to its relative weight. Denote by $w_i(j,t_1,t_2)$ the a-posteriori probability that the portion of the utterance x_i sampled at state j is $x^i_{t_1}, ..., x^i_{t_2}$. Applying Base rule yields :

$$w_i(j,t_1,t_2) = \frac{\sum f(x^i,s)}{f(x^i)}$$

where the sum is performed over all the state sequences such that the time interval of the sojourn in the state j is from t_1 to t_2. An extension of the Forward-Backward algorithm can be applied for efficient computation of $w_i(j,t_1,t_2)$. Computing the expressions $w_i(j,t_1,t_2)$ is actually the main part of the E-step of the Baum-Welch considered as a special case of the EM-algorithm. In case of multi-dimensional observations and diagonal matrix covariances, the weights $w_i(j,t_1,t_2)$ are computed for all the observations components together. Once the weights are computed, the estimation can be performed for each component separately. Denote the parameter set we want to estimate by θ. Denote the parameter set associated with the state j by $\theta_j = \{\mu_{j,a}, \sigma^2_{j,a}, \sigma^2_j\}$. The current estimate at the beginning of the EM iteration is denoted by $\tilde{\theta}$. The EM auxiliary function is :

$$Q(\theta,\tilde{\theta}) = E(\log f(x,s,\theta)|x,\tilde{\theta}) = \sum_i\sum_s f(s|x^i,\tilde{\theta})\log f(x^i,s,\theta)$$

where s is the hidden state sequence that was used to produce x^i. Differentiating the auxiliary function with respect to $\mu_{j,a}$ yields :

$$\frac{\partial Q(\theta,\tilde{\theta})}{\partial \mu_{j,a}} =$$

$$= \sum_i\sum_s f(s|x^i,\tilde{\theta})\frac{\partial}{\partial\mu_{j,a}}\log f(x_i,s,\theta)$$

$$= \sum_i\sum_{t_1<t_2} w_i(j,t_1,t_2)\frac{\partial}{\partial\mu_{j,a}}\log f(x^i_{t_1},...,x^i_{t_2}|\theta_j)$$

$$= \sum_i\sum_{t_1<t_2} w_i(j,t_1,t_2)\frac{1}{\Delta t\sigma^2_{j,a}+\sigma^2_j}\sum_{t=t_1}^{t_2}(x^i_t-\mu_{j,a})$$

Setting this partial derivative to zero yields the re-estimation formula for $\mu_{j,a}$.

$$\hat{\mu}_{j,a} = \frac{\sum_i\sum_{t_1<t_2} w_i(j,t_1,t_2)\frac{1}{\Delta t\sigma^2_{j,a}+\sigma^2_j}\sum_{t=t_1}^{t_2} x^i_t}{\sum_i\sum_{t_1<t_2} w_i(j,t_1,t_2)\frac{\Delta t}{\Delta t\sigma^2_{j,a}+\sigma^2_j}} \qquad (2)$$

This formula is not a valid re-estimation expression since the unknown parameters appear on both sides of the equation. For the other parameters, namely $\sigma_{j,a}$ and σ_j, a closed-form expression can not be obtained. Gales and Young [4] proposed to overcome this difficulty by using the approximation $\Delta t\sigma^2_{j,a} \gg \sigma^2_j$. Russell [7] used the joint probability of the observations and the optimal shift as the target function for the maximization problem. In this manner a closed-form expression can be obtained for all the parameters. He also proposed to substitute the current parameter values on the right hand side of the re-estimation equations. Digalakis et al. [2] have considered the static random model as a special case of the state space dynamic model. They suggested to solve the maximization problem we have in the M-step using an inner EM-algorithm. The unknown values of the shift random variable a are the missing data for that inner EM.

In the next section we shall suggest a modification of the static random segmental model that overcomes the problems mentioned above.

3. SCALING THE MODEL PARAMETERS

We now present a model which we have termed scaled static random segmental model. It is similar to Russell's model that was presented in the previous section, except that

$$a \sim N\left(0, \frac{\sigma^2_{j,a}}{\Delta t}\right)$$

where Δt is the segment length i.e. the duration of the sojourn in the state j. The scaled static model asserts that

the variance of the mean trajectory is inversely proportional to the segment length.

A closed-form expression for the PDF of the data segment associated with the state j in the scaled model can be obtained by substituting $\frac{\sigma^2_{j,a}}{\Delta t}$ in place of $\sigma^2_{j,a}$ in equation (1). The re-estimation expression for $\mu_{j,a}$ in the scaled model can be obtained by substituting $\frac{\sigma^2_{j,a}}{\Delta t}$ in place of $\sigma^2_{j,a}$ in (2), thus yielding:

$$\hat{\mu}_{j,a} = \frac{\sum_i \sum_{t_1<t_2} w_i(j,t_1,t_2) \sum_{t=t_1}^{t_2} x_t^i}{\sum_i \sum_{t_1<t_2} w_i(j,t_1,t_2)\Delta t} \qquad (3)$$

As can be seen from comparing expressions (2) and (3), 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 model 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.

Scaling the model also enables us to obtain closed-form re-estimation formulae for the other parameters. Denote :

$$\alpha_j = \sigma^2_{j,a} + \sigma^2_j$$

Setting the derivative of $Q(\theta,\tilde{\theta})$ with respect to $\sigma^2_{j,a}$ to zero, yields the following relation :

$$k = \frac{1}{\alpha_j} \sum_i \sum_{t_1<t_2} w_i(j,t_1,t_2)\Delta t (E(x_i,t_1,t_2) - \mu_{j,a})^2 \qquad (4)$$

where k is the number of utterances. Direct computation reveals :

$$\frac{\partial Q(\theta,\tilde{\theta})}{\partial \sigma^2_j} = -\frac{1}{2\sigma^2_j} \sum_i \sum_{t_1<t_2} w_i(j,t_1,t_2)(\Delta t - 1) \qquad (5)$$

$$+\frac{1}{2\sigma^4_j} \sum_i \sum_{t_1<t_2} w_i(j,t_1,t_2)\Delta t V(x_i,t_1,t_2)$$

$$-\frac{1}{2\alpha_j}(k - \frac{1}{\alpha_j} \sum_i \sum_{t_1<t_2} w_i(j,t_1,t_2)\Delta t (E(x_i,t_1,t_2) - \mu_{j,a})^2)$$

Substituting (4) in (5) yields the following Baum-Welch re-estimation equations :

$$\hat{\sigma}^2_j = \frac{\sum_i \sum_{t_1<t_2} w_i(j,t_1,t_2)\, \Delta t\, V(x_i,t_1,t_2)}{\sum_i \sum_{t_1<t_2} w_i(j,t_1,t_2)(\Delta t - 1)}$$

$$\hat{\sigma}^2_{j,a} = \frac{1}{k} \sum_i \sum_{t_1<t_2} w_i(j,t_1,t_2)\Delta t (E(x_i,t_1,t_2) - \hat{\mu}_{j,a})^2 - \hat{\sigma}^2_j$$

It should be noted that the estimate for $\sigma^2_{j,a}$ can be negative. However, it can be seen from the PDF of the scaled static model that the actual parameter is not $\sigma^2_{j,a}$ but $\alpha_j = \sigma^2_j + \sigma^2_{j,a}$, and the estimate of α_j is always non-negative.

4. LINEAR RANDOM MODEL

Deng et al. [1] 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 :

$$x_t = \mu_{j,a} + \mu_{j,b}(\frac{t-t_1}{t_2-t_1} - \frac{1}{2}) + \epsilon_t \qquad t = t_1, ..., t_2$$

such that t_1 is the time index of the transition into the state j and t_2 is the time index of the end of the sojourn in the state j. Holmes and Russell [5] presented a 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 model can be written as :

$$x_t = \mu_{j,a} + a + (\mu_{j,b} + b)(\frac{t-t_1}{t_2-t_1} - \frac{1}{2}) + \epsilon_t \qquad t = t_1, ..., t_2$$

where $\mu_{j,a}$ and $\mu_{j,b}$ are fixed parameters, a and b are independent normal random variables :

$$a \sim N(0, \sigma^2_{j,a}) \quad , \quad b \sim N(0, \sigma^2_{j,b})$$

and ϵ_t is a Gaussian white noise term, $\epsilon_t \sim N(0, \sigma^2_j)$.

We now present the scaled version for the linear random segmental model. 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 along the time axis.

From equation (2) it can be seen that the modeling problems in the static model are created by the fact that the contribution of each sample to the estimation of $\mu_{j,a}$ is weighted by the term $\frac{1}{\Delta t \sigma^2_{j,a} + \sigma^2_j}$ which depends on the duration of the segment corresponds to that sample. In a similar manner, in the linear model the contribution of each sample to the estimation of $\mu_{j,b}$ is weighted by the term $\frac{1}{F_b(\Delta t)\sigma^2_{j,b} + \sigma^2_j}$ such that :

$$F_b(\Delta t) = \frac{\Delta t(\Delta t+1)}{12(\Delta t-1)} = \sum_{t=t_1}^{t_2} (\frac{t-t_1}{t_2-t_1} - \frac{1}{2})$$

In the scaled linear model, therefore, the variances of a and b are functions of the segment duration as follows :

$$a \sim N\left(0, \frac{\sigma^2_{j,a}}{\Delta t}\right) \quad , \quad b \sim N\left(0, \frac{\sigma^2_{j,b}}{F_b(\Delta t)}\right)$$

where Δt is the segment length. It can be shown that using the scaled linear models enables us to obtain a closed-form solution for the maximization problem we have in the M-step of the Baum-Welch procedure. In other words, explicit re-estimation expressions can be obtained for all the parameters of the linear model. For example, the re-estimation formula for $\mu_{j,b}$ is :

$$\hat{\mu}_{j,b} = \frac{\sum_i \sum_{t_1<t_2} w_i(j,t_1,t_2) \sum_{t=t_1}^{t_2} x_t^i (\frac{t-t_1}{t_2-t_1} - \frac{1}{2})}{\sum_i \sum_{t_1<t_2} w_i(j,t_1,t_2) F_b(\Delta t)}$$

	m-n	p-t	s-ae	t-s
data-base size	3626	1194	862	863
deterministic static	56.0	63.5	78.2	68.1
unscaled static	49.4	61.0	80.3	69.3
scaled static	52.7	65.1	79.6	71.4
deterministic linear	57.9	67.6	79.9	66.7
unscaled linear	59.3	64.0	88.4	72.1
scaled linear	61.9	67.7	88.9	74.8

Table 1: Phoneme classification rate results

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 [3]. 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 chose for evaluation is phonetic classification. In classification the correct segmentation (phoneme 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 phoneme 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. Context dependent phonetic units were 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 models we implemented for evaluation were:

1. Standard Gaussian HMM.
2. Static random model [7].
3. Scaled static random model.
4. Linear mean trajectory segment model [1].
5. Linear random model [5].
6. Scaled linear random model.

In Table I we present recognition results for some frequently occurring triphone contexts. The first data row indicates the number of triphone occurrences for each context. Half of the occurrences were used to train each model. The other half was used to test the models. As can be seen, the scaled model outperforms the previously suggested non-scaled mode.

6. CONCLUSIONS

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 modeling. One desirable attribute of the scaled model is that it leads to a simple training algorithm. More precisely, given some training set, an exact Baum-Welch type algorithm can be employed. On the other hand, in the non-scaled model, either an iterative algorithm or an approximated target function are required to handle the maximization problem we have in the M-step of the Baum-Welch procedure.

Acknowledgment

We thank SRI international for providing the segmented data, obtained by using the DECIPHER recognition system.

7. REFERENCES

[1] L. Deng, M. Aksmanovic, D. Sun and J. Wu, "Speech recognition using hidden Markov models with polynomial regression functions as non stationary states", IEEE Trans. Speech Audio Processing, vol. 2, pp. 507-520, 1994.

[2] V. Digalakis, M. Ostendorf and J. R. Rohlicek, "A dynamical system approach to continuous speech recognition". Proc. Int. Conf. Acoust., Speech, Signal Processing, pp. 289-292, 1992.

[3] V. Digalakis, P. Monaco and H. Murveit, "Genones: generalized mixture tying in continuous hidden Markov model-based speech recognizers", IEEE Trans. Speech Audio Processing, vol. 4, pp. 281-289, 1996.

[4] M. Gales and S. J. Young, "The theory of segmental hidden Markov models", technical report, Cambridge, 1993.

[5] W. Holmes and M. Russell, "Speech recognition using a linear dynamic segmental HMMs", Proc. Eurospeech, pp. 1611-1614, 1995.

[6] M. Ostendorf, V. Digalakis and O. A. Kimball, "From HMMs to segmental models: a unified view of stochastic modeling for speech recognition", IEEE Trans. Speech and Audio Processing, vol. 4, pp. 360-377, 1996.

[7] M. Russell, "A segmental HMM for speech pattern modeling", Proc. Int. Conf. Acoust., Speech, Signal Processing, pp. 499-502, 1993.