PARAMETRIC TRAJECTORY MODELS FOR SPEECH RECOGNITION

Herbert Gish *Kenney Ng*

BBN Systems and Technologies
70 Fawcett Street 15/1c, Cambridge MA 02138 USA

ABSTRACT

The basic motivation for employing trajectory models for speech recognition is that sequences of speech features are statistically dependent and that the effective and efficient modeling of the speech process will incorporate this dependency. In our previous work [1] we presented an approach to modeling the speech process with trajectories. In this paper we continue our development of parametric trajectory models for speech recognition. We extend our models to include time-varying covariances and describe our approach for defining a metric between speech segments based on trajectory models; it is important in developing mixture models of trajectories.

1. INTRODUCTION

The motivation for much of the work on trajectory or segmental models is that conventional HMM's do not effectively exploit the time dependence of speech frames [1,2,3]. The polynomial, parametric trajectory model we employed in [1] to exploit the time dependency in the speech process had some shortcomings. In particular, it could not account for the change in variance of the trajectory as a function of time. That is, the model required a constant covariance function over the whole trajectory. The way we dealt with this limitation in [1] was to propose the mixture model for parametric trajectories. The mixture model of trajectories deals with the issue of trajectory variability implicitly by allowing more choices for the trajectories. Our description of the trajectory models did not include our methodology for measuring the distance between speech segments based on trajectory models, which is important for the development of mixture models. In the following we will describe our approach to measuring distances.

We will also present a new approach to trajectory modeling that now allows for a changing covariance structure as a function of the position along the trajectory. We will describe the algorithm for training such models and compare it to mixture modeling on a vowel recognition experiment.

2. BACKGROUND - THE CURRENT PARAMETRIC TRAJECTORY MODEL

The parametric trajectory model treats each speech unit being modeled as a curve (or collection of curves) in feature space, where the features typically are cepstra and their derivatives. The class of trajectories that we have thus far considered have been low degree polynomials, though our formulation does permit other classes of trajectory models.

For the parametric trajectory we model each feature dimension of a speech segment as

$$c(n) = \mu(n) + e(n) \qquad \text{for} \quad n = 1, \ldots, N \tag{1}$$

where $c(n)$ are the observed cepstral features in a segment of length N, $\mu(n)$ is the mean feature vector as a function of frame number and represents the dynamics of the features in the segment, and $e(n)$ is the residual error term which we assume to have a Gaussian distribution. In addition, the errors are assumed to be independent from frame to frame. The mean feature vector models that we consider in this paper will be at most a quadratic function of time, i.e.,

$$\begin{aligned} \mu(n) &= b_1 + b_2 n + b_3 n^2 \quad \text{for} \quad n = 1, \ldots, N \\ &= \underline{\mathbf{z}}' \cdot \underline{\mathbf{b}} \end{aligned} \tag{2}$$

where $\underline{\mathbf{z}}' = [1 \; n \; n^2]$ and $\underline{\mathbf{b}}' = [b_1 \; b_2 \; b_3]$. A primary assumption for this model is that the residual $e(n)$ is uncorrelated between any time instants.

Equation 2 is the trajectory for a single feature and a complete description of the model requires the joint distribution of all the features.

If we let $c_{n,i}$ denote the i^{th} feature at time n we can write

$$c_{n,i} = \beta_{1,i} + \beta_{2,i} n + \beta_{3,i} n^2 + e_{n,i} \tag{3}$$

where n takes on the the values $n = 1, \ldots, N$ and $i = 1, \ldots, D$ with D equal to the number of features. Although we have required the residuals, $e_{n,i}$ to be uncorrelated across time, we assume that, at each instant of time, they are D-dimensions Gaussian random variables with zero mean and covariance matrix Σ. This correlation is sometimes referred to as *contemporaneous correlation*. The requirement of constant covariance over time is a serious limitation of this model and we will later consider methods for overcoming this limitation.

Notwithstanding the constant covariance limitation the model does exploit the dependency of features in time through the trajectory. By not allowing a time varying covariance we are assuming that our uncertainty along the trajectory is time independent and this is not an entirely adequate assumption.

2.1. Estimation of the Model

Estimation of the model means estimation of the trajectory which mean estimation of weights $\beta_{\mathbf{i}} = (\beta_{1,i}, \beta_{2,i}, \beta_{3,i})$ and estimation of covariance matrix between the residuals, Σ. We first write the trajectory equation for each feature as

$$\mathbf{c_i} = \mathbf{Z}\beta_{\mathbf{i}} + \mathbf{e_i} \quad i = 1, \ldots, D \tag{4}$$

which is a vector representation of Equation 3 where $c_{n,i}$ is the observed feature observed at N time instants. $\mathbf{Z}$ is the design matrix which is determined by the nature of the trajectory and for our case it is a second degree polynomial and $e_{n,i}$ is the vector of N residuals for the observed feature.

In anticipation that in estimating a model we will be dealing with segments representing the same phonetic units that are of different duration, we will normalize all segments to be of unit length. This normalization is reflected in the design matrix. Below we consider estimating the model parameters using the normalized design matrix.

Expanding out Equation 4 for a quadratic trajectory model and a segment with N frames, we get:

$$\begin{bmatrix} c_{1,i} \\ c_{2,i} \\ \vdots \\ c_{N,i} \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ 1 & \frac{1}{N-1} & (\frac{1}{N-1})^2 \\ \vdots & \vdots & \vdots \\ 1 & 1 & 1 \end{bmatrix} \begin{bmatrix} \beta_{1,i} \\ \beta_{2,i} \\ \beta_{3,i} \end{bmatrix} + \begin{bmatrix} e_{1,i} \\ e_{2,i} \\ \vdots \\ e_{N,i} \end{bmatrix} \quad \text{for} \quad i = 1, \ldots, D \tag{5}$$

or

$$c_{n,i} = \beta_{1,i} + \beta_{2,i} \left(\tfrac{n-1}{N-1}\right) + \beta_{3,i} \left(\tfrac{n-1}{N-1}\right)^2 + e_{n,i} \quad \text{for} \quad n = 1, \ldots, N \quad \text{and} \quad i = 1, \ldots, D. \tag{6}$$

For each feature the Maximum Likelihood (ML) and linear least square estimates of the parameters are given by

$$\hat{\beta}_{\mathbf{i}} = [\mathbf{Z}^{'}\mathbf{Z}]^{-1}\mathbf{Z}^{'}\mathbf{c_i} \tag{7}$$

If we let $\mathbf{C}$ be a matrix whose i^{th} column is $\mathbf{c_i}$, $\mathbf{B}$ be a matrix whose i^{th} column is $\beta_{\mathbf{i}}$ and $\mathbf{E}$ whose i^{th} column is $\mathbf{e_i}$ we have the matrix equation for all the feature equations as given by Equation 4

$$\mathbf{C} = \mathbf{ZB} + \mathbf{E} \tag{8}$$

with the corresponding solution for the parameters given by

$$\hat{\mathbf{B}} = \left[\mathbf{Z}'\mathbf{Z}\right]^{-1}\mathbf{Z}'\mathbf{C} \tag{9}$$

Using the same matrix notation we can estimate the covariance matrix Σ from the estimated residuals, i.e.,

$$\hat{\Sigma} = \frac{\hat{\mathbf{E}}'\hat{\mathbf{E}}}{N} = \frac{\left(\mathbf{C} - \mathbf{Z}\hat{\mathbf{B}}\right)^{'}\left(\mathbf{C} - \mathbf{Z}\hat{\mathbf{B}}\right)}{N}. \tag{10}$$

2.2. Pooling the Data

In the estimation of a model for a trajectory for a phonetic unit we will have a collection of speech segments from which to create the model. As we have noted previously these segments will have different durations and to accommodate this variation we scale all segments to have unit length. Even with the scaling accommodating the different durations we are still faced with the equation for the trajectory for each of the segments having a different design matrix which we can denote by $\mathbf{Z_k}$, for the k^{th} segment.

We form the total observation matrix, the combined design matrix and total residual matrix,

$$\mathbf{C_T} = \begin{bmatrix} \mathbf{C_1} \\ \vdots \\ \mathbf{C_K} \end{bmatrix}, \mathbf{Z_T} = \begin{bmatrix} \mathbf{Z_1} \\ \vdots \\ \mathbf{Z_K} \end{bmatrix}, \mathbf{E_T} = \begin{bmatrix} \mathbf{E_1} \\ \vdots \\ \mathbf{E_K} \end{bmatrix} \tag{11}$$

respectively, where K is the total number of segments being pooled.

Analogous to Equation 8 we have

$$\mathbf{C_T} = \mathbf{Z_T B} + \mathbf{E_T}, \tag{12}$$

with the analogous solution for trajectory parameters being,

$$\hat{\mathbf{B}} = \left[\mathbf{Z_T}'\mathbf{Z_T}\right]^{-1}\mathbf{Z_T}'\mathbf{C_T}. \tag{13}$$

Using the represention for the matrices given in Equation 11 we obtain

$$\hat{\mathbf{B}} = \left[\sum_{k=1}^{K} \mathbf{Z}_k'\mathbf{Z}_k\right]^{-1}\left[\sum_{k=1}^{K} \mathbf{Z}_k'\mathbf{Z}_k\hat{\mathbf{B}}_k\right] \tag{14}$$

where $\hat{\mathbf{B}}_k$ is the estimate of the trajectory parameters obtained from the k^{th} segment and the pooled estimate is seen to be a weighted combination of the individual segment estimates. The estimate for the covariance becomes

$$\hat{\Sigma} = \frac{\sum_{k=1}^{K}\left(\mathbf{C}_k - \mathbf{Z}_k\hat{\mathbf{B}}\right)^{'}\left(\mathbf{C}_k - \mathbf{Z}_k\hat{\mathbf{B}}\right)}{\sum_{k=1}^{K} N_k} \tag{15}$$

2.3. Likelihood of a Segment

Being able to compute the likelihood of segment coming from a particular model is a primary goal of the modeling. Once an estimate has been established for a particular phonetic unit it can then be used to evaluate the likelihoods of speech segments of having been generated by the model. For example, let Σ_m and B_m be the trajectory model parameters for phonetic unit m, (which is estimated from pooled data as given above). Then the likelihood of a sequence of speech features (a segment) being generated by this model will depend on the segment via the estimate of trajectory parameters $\hat{\mathbf{B}}$, the estimate of the covariance matrix $\hat{\Sigma}$, and, N, the number of frames in the segment. For our Gaussian model the likelihood is given by:

$$\begin{aligned} L(\hat{\mathbf{B}}, \hat{\Sigma}|\mathbf{B}_m, \Sigma_m) \; = \; & \\ (2\pi)^{-\frac{DN}{2}} \; |\Sigma_m|^{-\frac{N}{2}} \cdot \exp\left(-\frac{N}{2}\mathrm{tr}\left[\Sigma_m^{-1}\hat{\Sigma}\right]\right) \cdot & \\ \exp\left(-\frac{1}{2}\mathrm{tr}\left[\mathbf{Z}(\hat{\mathbf{B}} - \mathbf{B}_m)\Sigma_m^{-1}(\hat{\mathbf{B}} - \mathbf{B}_m)'\mathbf{Z}'\right]\right). & \end{aligned} \tag{16}$$

The above expression shows that the likelihood is not simply a function of the likelihoods for for the trajectories of the individual features. The interaction between the trajectories for the individual features is caused by the contemporaneous correlation existing between the residuals associated with the different features.

3. DISTANCE BETWEEN SPEECH SEGMENTS BASED ON THE TRAJECTORY MODEL

A mixture model for trajectories is similar to the conventional use of Gaussian mixture models except that the mean

of each term in the mixture is a trajectory, such as is the case in Equation 16. The motivation for using mixtures is to obtain a better representation of the types of trajectories that can represent a phonetic unit.

We discussed the EM algorithm for training such a mixture in [1] however we did not discuss an important prerequisite for developing mixture models and that is developing a metric for distances between segments based on their trajectory parameter estimates. The metric that we employed is based on a generalized likelihood ratio approach that we have often used in developing metrics. See [4] for example.

The basic idea is that we consider the hypothesis that the observations associated with two segments were generated by the same trajectory model and compare it to the alternative hypothesis that they weren't generated by the same model. The hypotheses forms the basis for a generalized likelihood ratio test and the negative of the log likelihood ratio is used as the distance.

More specifically, given two speech segments, X ($N1$ frames long) and Y ($N2$ frames long), we have the following hypothesis test:

- H_o: the segments were generated by the same model, and
- H_1: the segments were generated by different models.

If we let λ denote the likelihood ratio, then

$$\lambda = \frac{L_o}{L_1}, \tag{17}$$

giving

$$\lambda = \frac{L(X;\hat{\mathbf{B}},\hat{\Sigma})\ L(Y;\hat{\mathbf{B}},\hat{\Sigma})}{L(X;\hat{\mathbf{B}}_1,\hat{\Sigma}_1)\ L(Y;\hat{\mathbf{B}}_2,\hat{\Sigma}_2)}, \tag{18}$$

where the hat denotes the ML estimate. Note the common parameters in the numerator.

Using Gaussian likelihood expressions in Equation 18 for the trajectory models and simplifying, we obtain:

$$\lambda = \frac{|\mathbf{S}_1|^{\frac{N_1}{2}}|\mathbf{S}_2|^{\frac{N_2}{2}}}{|\mathbf{S}|^{\frac{N}{2}}}, \tag{19}$$

where $N = N_1 + N_2$, $\mathbf{S}_1$ and $\mathbf{S}_2$ are the sample covariance matrices for segments X and Y respectively, and $\mathbf{S}$ is the sample covariance matrix for the joint segment model.

The sample covariance matrix for the joint segment model can be rewritten as a sum of two matrices as follows:

$$\mathbf{S} = \mathbf{W} + \mathbf{D} \tag{20}$$

where

$$\mathbf{W} = \frac{N_1}{N}\mathbf{S}_1 + \frac{N_2}{N}\mathbf{S}_2 \tag{21}$$

and

$$\mathbf{D} = \frac{(\mathbf{Z}_1\hat{\mathbf{B}}_1 - \mathbf{Z}_1\hat{\mathbf{B}})'(\mathbf{Z}_1\hat{\mathbf{B}}_1 - \mathbf{Z}_1\hat{\mathbf{B}})}{N} + \tag{22}$$

$$\frac{(\mathbf{Z}_2\hat{\mathbf{B}}_2 - \mathbf{Z}_2\hat{\mathbf{B}})'(\mathbf{Z}_2\hat{\mathbf{B}}_2 - \mathbf{Z}_2\hat{\mathbf{B}})}{N}. \tag{23}$$

Note that the $\mathbf{W}$ matrix is a weighted sum of the covariance matrices of the two separate segments, and the $\mathbf{D}$ matrix is composed of the deviations between the segment trajectories and the trajectory of the joint model.

From Equation 20, we can factor out the $\mathbf{W}$ matrix and obtain the following expression for the sample covariance of the joint model matrix and its determinant:

$$\mathbf{S} = \mathbf{W}\ (\mathbf{I} + \mathbf{W}^{-1}\mathbf{D}) \tag{24}$$

and

$$|\mathbf{S}| = |\mathbf{W}|\ |\mathbf{I} + \mathbf{W}^{-1}\mathbf{D}|. \tag{25}$$

Substituting Equation 25 into Equation 19, we obtain:

$$\lambda = \frac{|\mathbf{S}_1|^{\frac{N_1}{2}}|\mathbf{S}_2|^{\frac{N_2}{2}}}{|\mathbf{W}|^{\frac{N}{2}}} \times \frac{1}{|\mathbf{I} + \mathbf{W}^{-1}\mathbf{D}|^{\frac{N}{2}}}. \tag{26}$$

which can be written as

$$\lambda = \lambda_{COV}\lambda_{TRAJ} \tag{27}$$

where

$$\lambda_{COV} = \frac{|\mathbf{S}_1|^{\frac{N_1}{2}}|\mathbf{S}_2|^{\frac{N_2}{2}}}{|\mathbf{W}|^{\frac{N}{2}}} \tag{28}$$

and

$$\lambda_{TRAJ} = \frac{1}{|\mathbf{I} + \mathbf{W}^{-1}\mathbf{D}|^{\frac{N}{2}}}. \tag{29}$$

This factorization of the likelihood ratio into two terms corresponding to the "distances" between segments based on matching covariances of the residuals and trajectory parameters, respectively.

From these likelihoods, we obtain our "distances" between segments by taking the negative of their logarithms:

$$d_{COV} = -\log(\lambda_{COV}) \tag{30}$$
$$= \frac{N}{2}\log|\mathbf{W}| - \frac{N_1}{2}\log|\mathbf{S}_1| - \frac{N_2}{2}\log|\mathbf{S}_2|$$

and

$$d_{TRAJ} = -\log(\lambda_{TRAJ}) \tag{31}$$
$$= \frac{N}{2}\log|\mathbf{I} + \mathbf{W}^{-1}\mathbf{D}|.$$

Since the generalized likelihood ratio is always greater than zero and less than unity, the above "distances" are always positive, although they may not satisfy the triangle inequality. In our experiments we have found using the d_{TRAJ} distance measure preferable to using $d_{TRAJ} + d_{COV}$. A detailed discussion and analysis of these distance measures for the constant trajectory case under the assumption that the probability models are Gaussian can be found in [4].

4. TIME-VARYING COVARIANCE

We have already noted how the covariance of the residual being constant over time was fairly restrictive. In order go beyond these restrictions we base our approach on the generalized least squares approach (also ML) which includes temporal variation in covariance of the residuals. If we let $\mathbf{\Omega}$ denote the covariance matrix for the N residuals, the solution for the trajectory parameters, $\hat{\beta}_\mathbf{i}$, associated with feature i, (assuming Ω is known), is given by

$$\hat{\beta}_{\mathbf{i}} = [\mathbf{Z}^{'}\mathbf{\Omega}_{\mathbf{i}}^{-1}\mathbf{Z}]^{-1}\mathbf{Z}^{'}\mathbf{\Omega}_{\mathbf{i}}^{-1}\mathbf{c}_{\mathbf{i}}, \quad i = 1, \ldots, D \tag{32}$$

where $\mathbf{\Omega_i}$ is that part of $\mathbf{\Omega}$ relevant to the i^{th} feature. The only difficulty is that $\mathbf{\Omega}$ is not known. The approach that we followed was to first restrict the class of covariance matrices that we were interested in. Then we employed an iterative procedure for estimating $\mathbf{\Omega}$ and reestimating the model parameters. (Note that knowledge of the $\mathbf{\Omega}$ and the trajectory parameters of the model permits computation of the likelhoods of segments based on the Gaussian model since we then have a fully specified multivariate Gaussian model.)

We restricted the time-variation of the covariance to be limited to having three different covariance matrices existing over a segment i.e., we allowed a different covariance matrix for each third of a segment. The first step in the estimation procedure was to obtain parameter estimates from Equation 14, in order to build our initial models. Using the parameters obtained from this estimation process were then able to estimate the residuals at all times along the trajectory. The estimated residuals then permitted us to compute separate covariance matrices for each of the designated segments of the trajectory. This step provides us with our estimate $\hat{\mathbf{\Omega}}$ to be used in Equation 32. In this Equation $\hat{\mathbf{\Omega}}_{\mathbf{i}}$ will simply be a diagonal matrix with three different variances along the diagonal.

5. A VOWEL CLASSIFICATION EXPERIMENT

To evaluate the trajectory model, we performed experiments on a speaker independent vowel classification task. The corpus for this task consists of 16 vowels: 13 monothongs /iy, ih, ey, eh, ae, aa, ah, ao, ow, uw, uh, ux, er/ and 3 diphthongs /ay, oy, aw/. The vowels are excised, using the given phonetic segmentations, from the acoustically phonetically compact portion of the TIMIT corpus without any restrictions on the phonetic contexts of the vowels. From the 420 available speakers, 370 are used for training and the remaining 50 are used for testing. The test speakers are the same as those used in [5]. There is a total of 15,116 training tokens and 1,871 test tokens.

After the tokens are extracted, segment statistics are computed for each token several trajectory models are trained for each of the 16 vowels The models that we have trained and evaluated are

1. Gaussian with diagonal covariance matrix for the residuals
2. Gaussian with full covariance matrix for the residuals
3. Gaussian mixture model
4. Gaussian with the time-varying covariance.

Since the segment boundaries are known, the maximum *a posteriori* probability rule is used for classification of an unknown test segment k coming from model m:

$$\max_{m} \left[\mathrm{L}(\hat{\mathbf{B}}_k, \hat{\Sigma}_k | \mathbf{B}_m, \Sigma_m) \mathrm{p}(N|m) \mathrm{p}(m) \right] \tag{33}$$

where $\mathrm{p}(N|m)$ is the probability that phoneme m has length N, and is computed as a histogram of the training segment durations. In order to match the dynamic ranges of the likelihood term and $\mathrm{p}(N|m)$, an exponential weighting factor is placed on the duration term and selected to optimize performance on the training set.

Model	Diag-Cov	Full-Cov	Mixt.	Time-var.
Quadratic	61.94	64.29	65.74	66.06
Linear	61.09	63.65	63.81	62.10
Constant	57.83	61.67	61.89	58.52

Table 1. Percent correct performance of different trajectory models for different degree polynomials for vowel classification.

The results of our experiments are presented in Table 1. We note that the mixture model employed full covariance Gaussians and that the number of terms in each mixture varied, depending on the amount of data, but typically contained about eight terms. The clustering for initializing the mixture models was done by building and cutting dendrograms. We can see that the quadratic is best under all modeling situations and that the time-varying quadratic gave the best performance. Why the time-varying approach fared relatively poorly for the linear and constant trajectories needs futher investigation.

6. DISCUSSION

We have reviewed our approach to trajectory modeling and have presented a new way to generalize its capabilities. In particular we developed a method for modeling the time-varying covariances associated with a trajectory which reflects our uncertainty about the trajectory location. This approach was compared to our original model as well as trajectory mixture model. In addition we described our metric for measuring distance between trajectories.

We observe that the advanced methods that we have developed, mixture models and time-varying covariances, have the potential for being combined, i.e., having a Gaussian mixture models of trajectories in which the individual Gaussians have time-varying covariances.

REFERENCES

1. H. Gish and K. Ng, "A segmental speech model with applications to word spotting," in *Proc. ICASSP 1993*, pp. 447-450, 1993.
2. M. Ostendorf and S. Roukos, "A stochastic segment for phoneme-based continuous speech recognition," *IEEE Trans. on Acoust., Speech and Signal Proc.*, vol. 37, no. 12, pp. 1857-1869, 1989.
3. W. Goldenthal, "Statistical Trajectory Models for Phonetic Recognition," Ph.D Thesis, Massachusetts Institute of Technology, 1994.
4. H. Gish, M. H. Siu, and J. R. Rohlicek, "Segregation of Speakers for Speech Recognition and Speaker Identification" in *IEEE ICASSP* 1991, pp. 873-876.
5. H. M. Meng, V. W. Zue, and H. C. Leung, "Signal Representation, Attribute Extraction, and the Use of Distinctive Features for Phonetic Classification," in *Proc. DARPA Workshop on Speech and Natural Language*, Pacific Grove, CA, February, 1991, pp. 176-181.