STATISTICAL MODELING OF SPEECH FEATURE VECTOR TRAJECTORIES BASED ON A PIECEWISE CONTINUOUS MEAN PATH

Mark M. Thomson

Department of Electrical and Electronic Engineering
The University of Auckland
Private Bag 92019, Auckland, New Zealand

ABSTRACT

This paper presents a new statistical model of the trajectories of speech feature vectors. In this model each vector is assumed to correspond to a point on a mean path that consists of a number of concatenated straight line segments. The model characterizes both the deviation of the trajectory from the mean path and the deviation from the mean rate at which the vectors move through the vector space in a way that avoids the conditional independence assumption implicit in hidden Markov modeling.

The model is formulated using a state space approach in which the state vector consists of only two elements. These represent the position on the mean path corresponding to the present observation vector and the rate at which points on the mean path are moving through the vector space. A method for estimating the parameters of the model using the Expectation Maximization algorithm is presented.

1. INTRODUCTION

One of the key tasks in speech recognition based on statistical methods is the calculation of the class conditional probability density, $p(\mathbf{Y}_1^N \mid W)$, where $\mathbf{Y}_1^N = \{\mathbf{y}_1, \mathbf{y}_2, \cdots \mathbf{y}_N\}$ is an observed sequence of N feature vectors, and W is a particular acoustic unit, such as a word or a phoneme. Evaluation of $p(\mathbf{Y}_1^N \mid W)$ requires that a model exist of the statistical characteristics of feature vector sequences generated in uttering the acoustic unit W.

The most common approach is to use Hidden Markov Models (HMMs) in which each feature vector is assumed to be drawn from one of a finite number of probability distributions, each associated with a particular discrete state [1]. It is common in using this approach for recognition to use the joint likelihood $p(\mathbf{Y}_1^N S_1^N \mid W)$ as an approximation to $p(\mathbf{Y}_1^N \mid W)$, where S_1^N represents the most likely sequence of states. However a drawback of the HMM approach is that it assumes that observation vectors are statistically dependent on only the state in which they are produced, and not on previous observation vectors.

An different approach is the Stochastic Linear System Model (SLSM) in which the feature vector sequence is divided into segments, and the vectors within each segment are assumed to be noisy measurements of the response, $\mathbf{x}_k$, of some linear time-invariant system to a sequence of uncorrelated random vectors [2]. In this case, the state of the model is the continuous-valued vector, $\mathbf{x}_k$, rather than a discrete quantity as in HMMs.

However, in the SLSM approach, the feature vectors are statistically dependent on only the present state, as in HMMs. Also, it is necessary to map each feature vector into one of the time-invariant segments that make up the model. Because this is done using linear time warping, there is no explicit way to model dynamic variations in the rate at which feature vectors move along their trajectories.

In this paper a new method of modeling sequences of feature vectors is presented that does allow both the statistical dependence between observation vectors, and the rate at which vectors move along their trajectories to be characterized. The method is similar to the SLSM approach in that it makes use of a continuous state vector. However it also incorporates an explicit description of the mean vector trajectory, which is parameterized by representing it as a concatenation of straight line segments. The model permits the joint likelihood of the observation sequence and the most likely state vector sequence to be determined, in a way that is analogous to the calculation of $p(\mathbf{Y}_1^N S_1^N \mid W)$ in HMMs.

2. THE PIECEWISE CONTINUOUS MEAN PATH MODEL

A basic assumption underlying the model presented here is that for an ùtterance of a given acoustic unit

0-7803-2431-5/95 $4.00 © 1995 IEEE

each observed feature vector corresponds to some point on a mean path which the vectors take through the vector space. This can be regarded as equivalent to the assumption in a HMM representation that each vector is drawn from the distribution for a particular state and therefore has a particular mean value. We denote by $\mathbf{u}_k$ the point on the mean path corresponding to an observed vector $\mathbf{y}_k$.

A further assumption is that the mean path can be approximated by a concatenation of straight line segments. If the segments are numbered from 1 to P, the path can be represented by $P+1$ segment endpoint vectors, $\gamma_0, \gamma_1, \ldots \gamma_P$. In order to describe the rate at which the vectors $\mathbf{u}_k$ move along the mean path, the symbol d_k is used to represent the distance along the mean path from γ_0 to $\mathbf{u}_k$. If $\mathbf{u}_k$ lies on the ith segment, it can be expressed in terms of d_k by

$$\mathbf{u}_k = \gamma_i + (d_k - \zeta_i)\beta_i, \tag{1}$$

where

$$\zeta_i = \sum_{j=1}^{i} ||\gamma_j - \gamma_{j-1}||_2 \tag{2}$$

is the Euclidean distance along the mean path from γ_0 to γ_i, and

$$\beta_i = (\gamma_i - \gamma_{i-1}) / \, ||(\gamma_i - \gamma_{i-1})||_2 \tag{3}$$

is the unit length vector representing the direction of the ith segment.

The model describes two things: the way that $\mathbf{y}_k$ is related $\mathbf{u}_k$, and the way that $\mathbf{u}_k$ is related to $\mathbf{u}_{k-1}$. To model $\mathbf{y}_k$, we define

$$\mathbf{z}_k = \mathbf{y}_k - \mathbf{u}_k \tag{4}$$

and assume that $\mathbf{z}_k$ is a correlated Gaussian random vector. To describe the correlation in $\mathbf{z}_k$, we represent it by

$$\mathbf{z}_k = g\mathbf{z}_{k-1} + \mathbf{v}_{k-1}, \tag{5}$$

where g is a scalar between 0 and 1, and $\mathbf{v}_k$ is an uncorrelated Gaussian random vector with covariance $\mathbf{\Sigma}_v$. It is possible that $\mathbf{\Sigma}_v$ may depend on the segment i in which the vector $\mathbf{v}_k$ lies.

The rationale behind (5) is that we expect the rate at which the observed vectors, $\mathbf{y}_k$, move through the vector space in the direction of the mean path to be approximately the same as the rate at which the $\mathbf{u}_k$ vectors move along the mean path. This implies that on average $\mathbf{z}_k$ and $\mathbf{z}_{k-1}$ are parallel, so that $\mathbf{z}_{k-1}$ and the mean of $\mathbf{z}_k$ are related by a scalar.

To model the movement of the $\mathbf{u}_k$ vectors, we first define the variable c_k to be the rate at which these vectors move along the mean path. That is,

$$d_k = d_{k-1} + c_{k-1}. \tag{6}$$

Further, we assume that c_k is a correlated Gaussian random variable, with a non-zero mean value μ_c. We also assume that the correlation in c_k can be described by defining

$$e_k = c_k - \mu_c \tag{7}$$

and writing

$$e_k = he_{k-1} + f_{k-1}, \tag{8}$$

where h is a scalar between 0 and 1, and f_k is an uncorrelated Gaussian random variable with variance σ_f.

It is useful to put the model described by (1)-(8) in a standard state space form. To do this, we define the system state as

$$\mathbf{x}_k = \begin{bmatrix} d_k \\ e_k \end{bmatrix}, \tag{9}$$

from which (6)-(8) gives the state transition equation,

$$\begin{bmatrix} d_{k+1} \\ e_{k+1} \end{bmatrix} = \begin{bmatrix} 1 & 1 \\ 0 & h \end{bmatrix} \begin{bmatrix} d_k \\ e_k \end{bmatrix} + \begin{bmatrix} \mu_c \\ 0 \end{bmatrix} + \begin{bmatrix} 0 \\ f_k \end{bmatrix}, \tag{10}$$

In terms of the state vector, $\mathbf{y}_k$ can be represented by

$$\mathbf{y}_k = \begin{bmatrix} \beta_i & \mathbf{0} \end{bmatrix} \begin{bmatrix} d_k \\ e_k \end{bmatrix} + (\gamma_i - \zeta_i\beta_i) + \mathbf{z}_k, \tag{11}$$

where $\mathbf{0}$ is a column vector whose elements are all zero.

Note that if we regard $\mathbf{y}_k$ as the output of the system, then $\mathbf{z}_k$ in (11) can be regarded as measurement noise. However because $\mathbf{z}_k$ is a colored noise sequence, such a formulation would not be directly amenable to state estimation by Kalman filtering. To overcome this, a system with white observation noise can be produced using the method of measurement differencing [3]. To do this we define

$$\begin{aligned} \mathbf{n}_k &= \mathbf{y}_{k+1} - g\mathbf{y}_k \\ &= \mathbf{H}\mathbf{x}_k + \mu_c\beta_i + (1-g)(\gamma_i - \zeta_i\beta_i) + \mathbf{v}_k, \end{aligned} \tag{12}$$

where

$$\mathbf{H} = \begin{bmatrix} \beta_i & \mathbf{0} \end{bmatrix} \begin{bmatrix} 1 & 1 \\ 0 & h \end{bmatrix} - g \begin{bmatrix} \beta_i & \mathbf{0} \end{bmatrix}. \tag{13}$$

(12) can be regarded as the output equation since v_k is white.

3. PARAMETER ESTIMATION

In order to be able to use the model to estimate the likelihood of an observation sequence, it is necessary to first estimate the parameters of the model from a large number of reference utterances of each acoustic unit of interest. The vectors, $\gamma_0, \gamma_1, \ldots \gamma_P$ can be estimated as follows. For each reference sequence, we measure the total Euclidean distance along the path traced out by the sequence from one vector to another. Then we locate the $P+1$ points on each path that divide the trajectory into P equal length segments. Finally, we average the points corresponding to each breakpoint to produce the required estimates.

The remaining parameters, g, h, μ_c, σ_f and $\mathbf{\Sigma}_v$, can be estimated using the Expectation Maximization (EM) algorithm. Denoting by Θ the set of unknown parameters, we seek the values that maximize $p(\mathbf{Y} \mid \Theta)$, where $\mathbf{Y}$ is now the collection of all reference sequences. Using the EM algorithm [4], this is achieved by iteratively maximizing

$$Q(\Theta, \hat{\Theta}) = E\{\log p(\mathbf{Y}E \mid \Theta) \mid \mathbf{Y}\hat{\Theta}\} \tag{14}$$

with respect to Θ, where E represents the collection of sequences of values of e_k corresponding to $\mathbf{Y}$, and $\hat{\Theta}$ is an initial estimate of Θ. Note that a sequence of e_k values contains all the necessary information about the corresponding state vector sequence.

In a manner similar to that presented in [2], we proceed by maximizing $\log p(\mathbf{Y}E \mid \Theta)$ and then take the expected value of the result. Because $\mathbf{v}_k$ and f_k are assumed to be Gaussian, the objective function can be written as

$$\begin{aligned} J &= [\log p(\mathbf{Y} \mid E\Theta) + \log p(E \mid \Theta)] \\ &= -\frac{1}{2}\sum_k \{\log|\mathbf{\Sigma}_v| + (\mathbf{y}_k - \bar{\mathbf{y}}_k)^T \mathbf{\Sigma}_v^{-1}(\mathbf{y}_k - \bar{\mathbf{y}}_k)\} \\ &\quad -\frac{1}{2}\sum_k \{\log \sigma_f^2 + (e_k - \bar{e}_k)^2/\sigma_f^2\} \\ &\quad + \ constant \quad terms, \end{aligned} \tag{15}$$

where

$$\bar{\mathbf{y}}_k = \beta_i d_k + (\gamma_i - \zeta_i \beta_i) + \mathbf{z}_k, \tag{16}$$

$$\bar{e}_k = h e_{k-1}, \tag{17}$$

and the summation in (15) extends over all frames of all reference sequences.

Setting to zero the derivatives of J with respect to the unknown parameters produces expressions that involve the quantities e_k, e_k^2, $e_k e_{k-1}$, d_k, d_k^2, and $d_k d_{k-1}$. Updating the unknown parameters using the EM algorithm involves taking the expected values of these quantities, and these can be obtained using the Kalman smoothing algorithm as discussed in [2] and [5]. However it is important that the Kalman smoother be applied to the signal $\mathbf{n}_k$, rather than $\mathbf{y}_k$, in order to satisfy the assumption of white observation noise.

Note that in the smoothing process it is necessary to make use of initial values of $\mathbf{\Sigma}_v$ and σ_v^2 which depend on which straight line segment the vector $\mathbf{u}_k$ is located in. Although the true position of this vector is not known, we use the position estimated by the Kalman smoother to determine which of these variance values to use for each frame.

4. DISCUSSION

The continuous mean path model does not permit the true value of $p(\mathbf{Y}_1^N \mid W)$ to be easily calculated. However the joint likelihood of the observation sequence and the most likely mean path can be obtained as is often used in HMMs. The most likely mean path for a particlar observation sequence is simply that obtained from the Kalman smoother. The log of the required likelihood can then be computed from (15).

The model presented here offers a number of attractive features. Unlike the HMM and the SLSM it provides an effective means of describing correlation between adjacent feature vectors, while requiring a much smaller state vector than the SLSM. Two alternative ways of representing interframe correlation are the stochastic segment model [6] and the linear predictive HMM [7]. The stochastic segment model treats a complete segment as a multivariate Gaussian random variable characterised by a sequence of mean vectors and a very large covariance matrix. However the very large number of elements in the covariance matrix make training this model very difficult. In comparison with this approach, the model presented here has a very much smaller number of parameters.

The model presented here has similarities to the linear predictive HMM, but differs in the use of a continuous state vector. In addition, in the linear predictive HMM, interframe correlation is represented by assuming that the observation vectors are the output of a vector autoregressive system. In contrast in the model presented here, it is the difference between the observation vectors and the corresponding points on the mean path that are the output of an autoregressive system.

This paper has described the mathematical structure of the model based on a continuous mean path, and a method for estimating its parameters. It is hoped that the model will provide a tool for further investigations aimed at obtaining a better understanding of the nature of speech vector trajectories, with the ultimate goal of developing improved recognition techniques.

5. REFERENCES

[1] L. R. Rabiner, "A tutorial on hidden Markov models and selected applications in speech recognition," *Proc. IEEE*, vol. 77 no. 2, Feb. 1989.

[2] V. Digalakis, J. R. Rohlicek and M. Ostendorf, "ML estimation of a stochastic linear system with

the EM algorithm and its application to speech recognition," *IEEE Trans. Speech and Audio Processing*, vol. 1, no. 4, pp 431-442, Oct. 1993.

[3] A. E. Bryson and L. J. Henrikson, "Estimation using sampled data containing sequentially correlated noise," *J. Spacecr. Rockets*, vol. 5, pp 662-665, June 1968.

[4] A. P. Dempster, N. M. Laird and D. B. Rubin, "Maximum likelihood estimation from incomplete data," *J. Royal Stat. Soc. (B)*, vol. 39, no. 1, pp 1-38, 1977.

[5] H. E. Ruach, F. Tung and C. T. Striebel, "Maximum likelihood estimates of linear dynamic systems," *AIAA J.* , vol. 3, pp 1445-1450, Aug. 1965.

[6] M. Ostendorf and S. Roukos, "A stochastic segment model for phoneme-based continuous speech recognition," *IEEE Trans. Acoust. Speech, Signal Processing*, vol. 37, no. 12, pp 1857-1869, 1989.

[7] P. Kenny, M. Lennig, and P. Mermelstein, "A linear predictive HMM for vector-valued observations with applications to speech recognition," *IEEE Trans. Acoust. Speech, Signal Processing*, vol. 38, no. 2, pp 220-225, 1990.