CONTINUOUS SPEECH RECOGNITION USING A HIERARCHICAL BAYESIAN MODEL

Fériel MOURIA-BEJI[(1), (2)]
(1) *ENSI/LIA. Artificial Intelligence Group.*
ATCE, 3 Avenue Jean Jaurès, 1001 Tunis, Tunisia
E-mail : feriel.beji@intercom.tn
(2) *INRIA/LORIA. Villers–lès–Nancy, France*

ABSTRACT

This work proposes a stochastic model for continuous speech recognition that provides automatic segmentation of spoken utterances into phonemes and facilitates the quantitative assessment of uncertainty associated with the identified utterance features. The model is specified hierarchically within the Bayesian paradigm. At the lowest level of the hierarchy, a Gibbs distribution is used to specify a probability distribution on all the possible partitions of the utterance. The number of partitioning elements which are phonemes is not specified a priori. At higher level in the hierarchical specification, random variables representing phoneme durations and acoustic vector values are associated with each phoneme and frame. Estimation of the posterior distribution is done using Gibbs sampler scheme.

1. INTRODUCTION

By assuming the model parameters to be random variables, the posterior distribution of these parameters given the observed data can be constructed. This in turn makes inference about the model parameters more efficient. This work proposes a stochastic model for continuous speech recognition that provides automatic segmentation of spoken utterances into phonemes and facilitates the quantitative assessment of uncertainty associated with the identified utterance features [1]. Through techniques like Gibbs sampling [2], this parametrization provides a mechanism for estimating the posterior distribution of the spoken utterance. Posterior distribution models provide a broader and potentially more powerful class of discriminant functions [3]. They need not be resticted to fixed length features different versions of which are used and which are problematic from a statistical point of view. Posterior distributions have been used successfully in HMM's. For segmental posterior distributions, however, somme additional difficulties are encountred. The bigger problem relates to conditional independence assumptions, which are theoretically inconsistent in some of the currently proposed segmental posterior distribution models. Successive triphones necessarily depend on each other. It is not reasonable to assume that a given phoneme is independent of all the feature vectors to which it does not correspond. The model is specified hierarchically within the Bayesian paradigm [4]. At the lowest level of the hierarchy, a Gibbs distribution is used to specify a probability distribution on all the possible partitions of the utterance. The number of partitioning elements which are phonemes is not specified a priori. In the second level of the model a random variable representing phoneme durations is associated with each phoneme. Partitioning segment durations are assumed to be drawn from a distribution centered around phoneme durations. This is a typical model for an empirical Bayes approach [5]. At higher level in the hierarchical specification, multinormal random variables representing acoustic values are associated with frames. Gibbs model is extensively used in the analysis of noisy images [1,6,7]. The attractiveness of these distributions originates from their Markovian property: Gibbs distributions are special cases of Markov random fields. Because of this property, the prior distribution can be specified in terms of local conditional distributions involving only nearby phonemes. As demonstrated by Geman and Geman this Markovian property also facilitates both sampling from the posterior distribution using a technique known as the Gibbs sampler and maximization of the posterior distribution using a method called simulated annealing. A final point concerning the hierarchical specification of the model is that observations at frames are considered to be independent given their phoneme associations.

2. MODEL DESCRIPTION

In speech recognition, a spoken utterance can be represented by a sequence (in time) of m symbols: $\mathbf{x} = (x_1, x_2, \ldots, x_m)$, where each symbol x_i can be one of c

phonemes (or other speech units), labelled $1, 2, \ldots, c$ with c finite. After the utterance speech signal has been processed, each phoneme gives rise to a set of varying number of p dimensional acoustic vectors. Thus, the problem considered can be formulated as follows. Given an observed time series $\{\mathbf{y}_t : t = 1, \ldots, T\}$ of p dimensional acoustic vectors, partition the time index set $S = \{t : t = 1, \ldots, T\}$ into subsets $S_1 = \{t : t = 1, \ldots, t_1\}$, $, \ldots, S_m = \{t : t = t_{m-1}, \ldots, T\}$ and identify the phoneme x_i generating the subset of vectors with index in S_i, $i = 1, \ldots, m$. We write $\mathbf{x}^*$ for the true but unknown sequence of phonemes and interpret this as a particular realization of a random vector $\mathbf{X} = (X_1, X_2, \ldots, X_m)$ where X_i assigns symbol to position i of the utterance. Similarly, each $\mathbf{y}_t$ is a particular realization of the random vector $\mathbf{Y}_t$. It is clear that the change-points $t_1, t_2, \ldots, t_{m-1}$ and the number of phonemes in the utterance m are values of random variables. Thus, we set $\mathbf{d} = (d_1, d_2, \ldots, d_m)$, where $d_i = (t_i - t_{i-1})$, $i = 1, \ldots, m$, with $t_0 = 1$ and $t_m = T$.

Using $f(\cdot)$ to denote probability density functions (p.d.f) and p to denote discrete probabilities of named events, we make the following assumptions:

The first-stage of the model supposes that the true sequence of phonemes $\mathbf{x}^*$ is a realization of a locally dependent Markov random field (M.r.f.) [8]. We shall be concerned with a first order M.r.f. in which the phoneme symbol at position i of the utterance string depends only on the adjacent phoneme symbols at position $(i-1)$ and $(i+1)$. For any sequence (r, s) of two consecutive phonemes, let:

$$U_k(r,s) = \begin{cases} 1 & if\ r = k \\ 0 & if\ r \neq k \end{cases} \quad V_l(r,s) = \begin{cases} 1 & if\ s = l \\ 0 & if\ s \neq l \end{cases}$$

and $\rho(U_k, V_l)$ the correlation coefficient between U_k and V_l. When $\rho = 1$ the two phonemes k and l are not associated. When $\rho = 1$ phoneme k always precedes phoneme l, and phoneme l always succeeds to phoneme k, with a similar interpretation for $\rho = -1$. The prior probability of $(M, \mathbf{X})$ is taken to be:

$$p(m, \mathbf{x} \mid \rho) \propto$$

$$\exp\left(-wm^2 + \sum_{i=2}^{m} \rho\left(U_{x_{i-1}}, V_{x_i}\right) + \sum_{i=1}^{m-1} \rho\left(U_{x_i}, V_{x_{i+1}}\right)\right) \quad (1)$$

where the potential wm^2 is used to discourage configurations having large numbers of phoneme symbols and ρ denotes the set of all correlation coefficients $\rho(U_k, V_l)$, $k, l = 1, \ldots, c$.

In the second stage of the model, a distribution for the mean time duration of each phoneme is specified. We shall assume that D_i, the number of acoustic vectors with time index in S_i, has a Poisson distribution with parameter λ_i that is $D_i \mid \lambda_i \mathcal{P}(\lambda_i)$ and that given $x_i = k$, λ_i has a Gamma distribution with parameters α_k and β_k that is $\lambda_i \mid x_i = k \to \mathcal{G}(\alpha_k, \beta_k)$ with density function

$$f\ (\lambda_i \mid \alpha_k, \beta_k) = \frac{\beta^\alpha}{\Gamma(\alpha)} \lambda_i^{\alpha-1} \exp(-\beta \lambda_i)$$

where $i = 1, \ldots, m$ and $k = 1, \ldots, c$. It follows that

$$f(d_i, \lambda_i \mid \alpha_k, \beta_k) = p(d_i \mid \lambda_i) \times f(\lambda_i \mid \alpha_k, \beta_k) \quad (2)$$

Also, the marginal distribution of D_i when the Poisson distribution is compounded with the Gamma distribution is the Negative Binomial distribution with parameters α_k and $1/(1 + \beta_k)$

$$p(d_i \mid x_i = k, \alpha_k, \beta_k)$$

$$= \frac{\Gamma(\alpha_k + d_i)}{\Gamma(\alpha_k)\, d_i!} \left(\frac{\beta_k}{1 + \beta_k}\right)^{\alpha_k} \left(\frac{1}{1 + \beta_k}\right)^{d_i} \quad (3)$$

In the sequel, we shall let $\mathbf{\Lambda} = \{\lambda_i : i = 1, \ldots, m\}$.

In the third and last stage of the model, we specify the mean distributions of the acoustic vectors $\mathbf{Y}_t$ for t in S_i. We shall assume that:

1. Given $\mathbf{x}$ the random vectors $\mathbf{Y}_t$, $t = 1, \ldots, T$ are conditionally independent i.e. $f(\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_T \mid \mathbf{x}) = \prod_{t=1}^{T} f(\mathbf{y}_t \mid \mathbf{x})$.
2. For $t \in S_i$, the distribution of $\mathbf{Y}_t$ depends only on x_i i.e. $f(\mathbf{y}_t \mid \mathbf{x}) = f(\mathbf{y}_t \mid x_i)$.
3. $\mathbf{Y}_t$ has a multivariate Normal distribution with mean $\mathbf{\Theta}_t$ and precision matrix $\mathbf{V}_t$ that is $\mathbf{Y}_t \mid (\mathbf{\Theta}_t, \mathbf{V}_t) \to \mathcal{N}_p(\mathbf{\Theta}_t, \mathbf{V}_t)$. Moreover, for $t \in S_i$ and $x_i = k$, $\mathbf{V}_t = \mathbf{V}_k$ a known symmetric positive definite precision matrix and the prior distribution of $\mathbf{\Theta}_t$ is multivariate Normal with mean vector $\mu_k \in \mathcal{R}^p$ and known symmetric positive definite precision matrix $\mathbf{\Sigma}_k$, that is $\mathbf{\Theta}_t \mid (x_i = k, \mu_k, \mathbf{\Sigma}_k) \to \mathcal{N}_p(\mu_k, \mathbf{\Sigma}_k)$. Then the posterior distribution of $\mathbf{\Theta}_t$ when $\mathbf{Y}_t = \mathbf{y}_t$, $(t \in S_i)$ is a multivariate Normal distribution [9,10], with mean vector

$$\mu_k^* = (\mathbf{\Sigma}_k + \mathbf{V}_k)^{-1} (\mathbf{\Sigma}_k \mu_k + \mathbf{V}_k \mathbf{y}_t)$$

and precision matrix

$$\mathbf{\Sigma}_k^* = (\mathbf{\Sigma}_k + \mathbf{V}_k)$$

that is the posterior distribution of $\mathbf{\Theta}_t$ given $\mathbf{Y}_t = \mathbf{y}_t$ $t \in S_i$ $x_i = k$ is

$$f(\mathbf{\Theta}_t \mid \mathbf{y}_t, \mu_k, \mathbf{V}_k, \mathbf{\Sigma}_k, t \in S_i,\ x_i = k)$$

$$\propto \exp\left\{(\theta_t - \mu_k^*)' \, \mathbf{\Sigma}_k^* \, (\theta_t - \mu_k^*)\right\} \quad (4)$$

From Bayes' rule, one has the following joint posterior density for the unknown parameters $m, \mathbf{x}, \mathbf{d}, \mathbf{\Lambda}$ and $\mathbf{\Theta} = \{\mathbf{\Theta}_t : t = 1, \ldots, T\}$

$$p(m, \mathbf{x}, \mathbf{d}, \mathbf{\Lambda}, \mathbf{\Theta} \mid \mathbf{y}) \propto \underbrace{p(\mathbf{x}, m)}_{1^{st}\ stage} \times \underbrace{p(\mathbf{d}, \mathbf{\Lambda} \mid \mathbf{x}, m)}_{2^{nd}\ stage}$$

$$\times \underbrace{f(\mathbf{y}, \mathbf{\Theta} \mid m, \mathbf{x}, \mathbf{d}, \mathbf{\Lambda})}_{3^{rd}\ stage} \quad (5)$$

where $p(\mathbf{x},m)$ is given by Eq.(1), $p(\mathbf{d},\mathbf{\Lambda} \mid \mathbf{x},m)$ is given by Eq.(2) and

$$p(\mathbf{d},,\mathbf{\Lambda} \mid \mathbf{x},m) = \prod_{i=1}^{m} f\left(d_i, \lambda_i \mid \alpha_k, \beta_k\right)$$

3. ESTIMATION

Estimation techniques that attempt to maximize the posterior distribution given by Eq.(5) may not adequately represent all plausible phoneme sequences for the given utterance. Also, due to random variation in acoustic data, there is often substantial uncertainty regarding the precise location of segment endpoints. Therefore, the posterior distribution is likely to be multimodal, with distinct modes corresponding to plausible positions of the change points. Thus estimation strategies that describe posterior uncertainty in utterance features are required. One possibility for representing such uncertainty is to generate samples from the posterior distribution of the utterance. A mechanism that can be used to generate such samples is Gibbs Sampler proposed in the imaging context by Geman and Geman. The basic requirement for implementing the Gibbs sampler is that full or reduced conditional distributions for all unknown quantities be available. In the hierarchical specification of our model, the full conditional distributions for the duration and acoustic related parameters given the sequence of phonemes may be derived using standard results in Bayesian inference [11,12]. For instance, sampling from the duration distribution of phoneme k can de done using Eq.(3).

Estimation of $m, \rho\left(U_k, V_l\right), \alpha_k, \beta_k, \mu_k, \mathbf{V}_k$ and $\mathbf{\Sigma}_k$, $k,l = 1,\ldots,c$, can be done from a training corpus. $\widehat{m}$ can be taken as the average number of phonemes per utterance in the training corpus. Estimation of $\rho\left(U_k, V_l\right)$, $k,l = 1,\ldots,c$ is as follows.

Let $\mathbf{N}$ be the $(c \times c)$ contingency table where cell (k,l) contains n_{kl} the number of times phoneme l follows phoneme k, $\quad k,l = 1,\ldots c$. Let:

$$n_{k+} = \sum_{l} n_{kl}, \quad n_{+l} = \sum_{k} n_{kl},$$

$$n_{\bar{k}+} = \sum_{k' \neq k} \sum_{l} n_{k'l}, \quad n_{k\bar{l}} = \sum_{k} \sum_{l' \neq l} n_{kl'}$$

then $\rho(k,l)$ is estimated by:

$$\widehat{\rho}(k,l) = \frac{n_{kl} - n_{k+}n_{+l}}{\sqrt{n_{k+}n_{\bar{k}+}n_{+l}n_{+\bar{l}}}}$$

this is the maximum likelihood estimate under the multinomial sampling model.

Estimation of $\mu_k, \mathbf{V}_k$ and $\mathbf{\Sigma}_k$ are obtained as follows:

$$\hat{\mu}_k = \frac{1}{\sum_j n_{jk}} \sum_{i,j} \mathbf{y}_{ijk}$$

$f(\mathbf{y}, \mathbf{\Theta} \mid m, \mathbf{x}, \mathbf{d}, \mathbf{\Lambda})$ is given by Eq.(4) and

$$f(\mathbf{y}, \mathbf{\Theta} \mid m, \mathbf{x}, \mathbf{d}, \mathbf{\Lambda}) \propto f(\mathbf{\Theta} \mid \mathbf{y}, m, \mathbf{x}, \mathbf{d}, \mathbf{\Lambda})$$

$$= \prod_{i=1}^{m} \prod_{t=t_{i-1}+1}^{t_i} f\left(\mathbf{\Theta}_t \mid \mathbf{y}_t, \mu_k, \mathbf{V}_k, \mathbf{\Sigma}_k, t \in S_i,\ x_i = k\right)$$

$$\widehat{\mathbf{V}}_k^{-1} = \frac{1}{\left(\sum_j n_{jk} - n_k\right)} \sum_{j=1}^{n_k} \sum_{i=1}^{n_{jk}} \left(\hat{\mu}_{jk} - \hat{\mu}_k\right)\left(\hat{\mu}_{jk} - \hat{\mu}_k\right)'$$

$$\hat{\mathbf{\Sigma}}_k^{-1} = \frac{1}{\left(n_k - 1\right)_k} \Sigma_{j=1}^{n_k} \left(\hat{\mu}_{jk} - \hat{\mu}_k\right)\left(\hat{\mu}_{jk} - \hat{\mu}_k\right)'$$

where $\mathbf{y}_{ijk}$ denotes the i^{th} acoustic vector in the j^{th} occurrence of phoneme k in the training corpus $k = 1,\ldots,c$, n_{jk} is the number of acoustic vectors in the j^{th} occurrence of phoneme k and n_k the number of times phoneme k is present in the training corpus.

An initial step in estimating the posterior distribution of the utterance is to identify plausible values of the hyperparameter w. Unfortunately, the intractability of the normalization constant prevents an analytic approach from being taken in the selection of this hypermarameter. The Gibbs sampler can be used to obtain samples from the prior, and examination of these samples provides guidance into appropriate choices for this hyperparameter. We begin by selecting values for $x_1,\ldots,x_m$. Then, using equation (3) we draw values for $d_1,\ldots,d_m$. Equation (4) generates values for $\theta_1,\ldots,\theta_T$. All these values are replaced in equation (5) from which new values of $x_1,\ldots,x_m$ are obtained. This completes one cycle of the Gibbs sampling scheme. Unfortunately, there is no criteria to determine how many cycles are needed for convergence.

4. EXPERIMENTAL RESULTS

The validation experiments were based on task-independent acoustic training, i.e., the vocabulary of the training text has been designed to have little coverage over that of recognition text. The speech data base is in Arabic. The training utterances consist of 80 phonetically rich sentences. The testing utterances consist of 300 sentences with 1481 words. All utterances were sampled from information requests of travel agencies. Speech was sampled at 16 kHz, blocked each 10 ms with a 25.6-ms window and parametrized using 17 Mel-frequency cepstral coefficient (MFCC) (energy included) unless otherwise specified. Utterances were recorded in two sessions over several days. Seven speakers were recorded. Four of them were used for development and the remaining for evaluation. All tests are in speaker-dependent mode. No speaker selection was performed to maximize the recognition rate.

Thirty-eight context-independent phone models including one silence model were used for all experiments. The

Speaker	%Cor	%Acc	Cor	Del	Sub	Ins	W
abj	98.99	98.92	1467	1	14	1	1481
abn	99.46	99.33	1474	0	8	2	1481
jar	98.92	98.65	1466	0	16	4	1481
fab	98.79	98.58	1464	0	18	3	1481
overall	99.04	98.87	5871	1	56	10	5924

Table 1: Word Recognition Results for different Speakers

initial segmentation of the training utterances were provided by an automatic time-alignment procedure. Typically, there were about 2753 segments per speaker in the training database.

The recognition task is described by a finite-state network with a vocabulary of 1770 words. The equivalent word-pair perplexity is 40. In our experiments, word transition probabilities are not used, i.e., the probabilities of all transitions from a node are equal.

To find the best sentence, the recognition system performs beam search with N-best sentence as final result [13]. If not otherwise stated, the beam size is fixed to 1000 and, for each utterance to be recognized, the system outputs an average of 833 complete sentences. In counting errors, only the top sentence is used.

We sampled the partition for a variety of hyperparameter values and found configurations using $w = 0.13$ and $m = 27$ to be appropriate. To speed the convergence of the algorithm, initial estimate of the phoneme sequences were obtained using Viterbi algorithm. Following initialization conditional distributions described in equations (1)-(3) were successively used to update phonemes, mean durations and mean acoustic vectors.

HTK toolkits were used for scoring the recognition results given in table 1 where Cor, Del, Ins, Sub, and W are, respectively, the number of correct words, deletions, insertions, substitutions and total number of words in the test speech, with $Cor = W - Del - Sub$. The accuracy is given by the ratio $(Cor - Ins)/W$.

5. CONCLUSION

This work described a stochastic model for continuous speech recognition. The proposed model is parametrized so that estimation of the posterior distribution is provided. A Gibbs distribution is used to specify a probability distribution on the space of all possible phoneme sequences of a spoken utterance. The attractiveness of this distribution originates from their Markovian property: Gibbs distributions are special cases of Markov random fields. Because of this property, the prior distribution of the true utterance sequence of phonemes can be specified in terms of local conditional distributions involving only neighboring phonemes. Estimation of the posterior distribution is done iteratively using Gibbs sampler scheme.

REFERENCES

[1] V. E. Johnson, “A Model for Segmentation and Analysis of Noisy Images”, Journal of the American Statistical Association, vol. 89, no. 425, pp. 230-241, 1994.

[2] S. Geman and D. Geman, “Stochastic Relaxation, Gibbs Distributions, and the Bayesian Restoration of Images”, IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 6, pp. 721-741, 1984.

[3] M. Ostendorf, V. V. Digalakis and O. A. Kimball, “From HMM to Segment Models: A Unified View of Stochastic Modeling for Speech Recognition”, IEEE Trans. on Speech and Audio Processing, vol. 4, no. 2, pp. 360-378, 1996.

[4] F. Mouria-Beji and J. P. Haton, “A hierarchical Bayesian model for continuous speech recognition”, To appear in Pattern Recognition Letters, PATREC-1560, Elsevier.

[5] J. J. Deely and D. V. Lindley, “Bayes Empirical Bayes”, Journal of the American Statistical Association, vol. 76, no. 376, pp. 833-841, 1981.

[6] J. E.Besag, “On the Statistical Analysis of Dirty Pictures”, Journal of the Royal Statistical Society, ser. B., vol. 48, no. 3, pp. 259-302, 1986.

[7] J.Marroquin, S. Mitter and T. Poggio, “Probabilistic Solution of Ill-Posed Problems in Computational Vision”, Journal of the American Statistical Association, vol. 82, no. pp. 397, 76-89, 1987.

[8] F. Mouria-Beji, “Segmental phoneme recognition using Markov random fields”, Submitted to IEEE Trans. on Speech and Audio Processing, Re: SAP-820.

[9] M. H. De Groot, “Optimal Statistical decisions”, McGraw-Hill, 1970.

[10] G. E. P. Box and G. C. Tiao, “Baysian Inference in Statistical Analysis”, Reading, MA: Addison-Wesley, 1973.

[11] A. E. Gelfand and Q. F. M. Smith, “Sampling Based Approach to Calculating Marginal Densities”, Journal of the American Statistical Association, vol. 85, no. 410, pp. 398-409, 1990.

[12] M. A. Tanner and W. H. Wong, “The Calculation of Posterior Distributions by Data Augmentation”, Journal of the American Statistical Association, vol. 82, no. 398, pp. 528-550, 1987.

[13] Y. Gong, J. P. Haton and F. Mouria-Beji, “Continuous Speech Recognition Based on High Plausibility Regions”, IEEE International Conference on Acoustics, Speech and Signal Processing, vol. 1, pp. 725-728, Toronto, Canada, 1991.