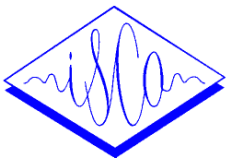

4th European Conference on
Speech Communication and Technology
EUROSPEECH '95
Madrid, Spain, September 18-21, 1995

ON THE USE OF BI-DIRECTIONAL CONTEXTUAL DEPENDENCE IN ACOUSTIC MODELING FOR SPEECH RECOGNITION

Qiang Huo[†‡] and Chorkin Chan[†]

[†]Department of Computer Science, The University of Hong Kong, Pokfulam Road, Hong Kong
[‡]ATR Interpreting Telecommunications Research Labs., 2-2 Hikaridai, Seika-cho Soraku-gun, Kyoto 619-02, Japan

ABSTRACT

With the motivation of utilizing bi-directional contextual dependence in acoustic modeling, in this paper, a bi-directional hidden Markov modeling approach for speech recognition is studied and the importance of bi-directional contextual dependence for speech recognition is identified by a series of comparative experiments. Furthermore, hidden Markov random field based acoustic modeling techniques using our previously proposed contextual vector quantization method and iterated conditional modes algorithm which is very suitable for the parallel processing implementation are also attempted. Their viability is confirmed by a series of preliminary experiments in a speaker independent isolated English letter recognition task.

10.21437/Eurospeech.1995-135

1. INTRODUCTION

Although hidden Markov modeling (HMM) techniques are currently among the most successful approaches to acoustic modeling for speech recognition, their performance is limited, partly because observation feature vectors at different times are assumed conditionally independent given the underlying state sequence and partly because the first-order Markov chain assumption on the hidden state sequence does not adequately model the temporal structure. In the past few years, much efforts have been made to study the correlation between neighboring feature vectors and different degrees of success have been reported. As for the Markov chain assumption, some efforts have been reported to extend the first-order assumption to second order one. There are also efforts to use Markov random field (MRF) model to replace the Markov chain assumption. Although the use of MRFs has a long history in image processing applications, only recently MRF theory began to attract some researchers' interest in speech recognition applications [6, 3, 5, 4].

Some cognitive experiments show that in a speech utterance, the appearance of a particular phonetic segment may constrain the range of phonemes that might appear next. Thus hearing a particular phoneme might enable the perceiver to shrewdly predict the phonemes likely to follow, narrowing the range of the phoneme candidates that would have to be considered. On the other hand, there are also evidences to show that each phoneme contains acoustic clues about the succeeding phoneme(s) derived from the alteration in its own pronunciation. We believe that this kind of bi-directional contextual dependence information also exists in a lower level and should be considered in acoustic modeling for speech recognition. This motivates the work presented in this paper.

2. STATISTICAL ACOUSTIC MODELING

In acoustic modeling for speech recognition, a speech utterance is usually represented by a sequence of observed feature vectors $\mathbf{X} = (\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_T)$. Each "observed data" $\mathbf{x}_t$, to be modeled as a realization of a random vector, can be interpreted as a partly observed version of a "complete data" $\mathbf{y}_t = (\mathbf{x}_t, z_t)$, where z_t is the missing value (hidden state) and takes one of a finite set of K qualitative values $\mathcal{G} = \{G_1, G_2, \cdots, G_K\}$ to represent a particular articulatory configuration. The "missing data" (hidden state) sequence $\mathbf{Z} = (z_1, z_2, \cdots, z_T)$ represents the corresponding temporal changes of the articulatory configurations. The probability density function (PDF) of $\mathbf{y}_t$ is assumed to be

$$p(\mathbf{y}_t) = p(\mathbf{x}_t, z_t) = p(\mathbf{x}_t|z_t) \cdot Pr(z_t) \ . \qquad (1)$$

Then in terms of the factorization in equation (1), the marginal density for $\mathbf{x}_t$ is

$$p(\mathbf{x}_t) = \sum_{j=1}^{K} p(\mathbf{x}_t|z_t = G_j) \cdot Pr(z_t = G_j) \ . \qquad (2)$$

In this case, $\mathbf{x}_t$ is said to have a *finite mixture distribution*. Under different assumptions for the dependence structure associated with the random variables $\mathbf{Y} = (\mathbf{X}, \mathbf{Z})$, different acoustic models can be identified. For example, we assume the $\mathbf{x}_t$'s are independent given z_t's. When the z_t's are assumed to be independent and thus marginally the z_t's are samples from a multinomial model, the mixture model is then considered as a hidden multinomial model. When the z_t's are assumed to follow a Markov chain on the state space $\mathcal{G}$, this lead to HMM which has become a predominant approach for speech recognition during the last decade. A more general approach is to assume the z_t's following a MRF and this lead to the so called hidden MRF model.

3. BI-DIRECTIONAL HMM OF SPEECH

In hidden Markov modeling (HMM) of speech, the hidden state sequence $z_1, z_2, \cdots, z_T$ are arbitrarily assumed to be a partial observation of a Markov chain. According to the property of a Markov chain, a Markov chain looked at in reverse order will be a Markov process, but in general its transition probabilities will depend on time and hence it will not be a Markov chain. If one pretends to be blind of the success of the conventional HMM where $z_1, z_2, \cdots, z_T$ are arbitrarily assumed to follow a Markov chain, on the contrary, in view of the fact of the bi-directional contextual dependent nature in the speech production process as stated above, one may naturally

Table 1: Performance comparison (% correct) between forward (fwd) and backward (bwd) HMMs

type	Type of HMMs					
	DHMM		EG-CDHMM		LR-CDHMM	
	close	open	close	open	close	open
fwd	85.0	73.9	66.9	65.1	71.8	68.6
bwd	82.5	69.9	68.0	66.5	71.9	68.4

ask the following question: how about the performance of the backward HMM where z_T, z_{T-1}, $\cdots$, z_1 are also assumed to be a partial observation of a Markov chain? In order to answer this question, a series of experiments are conducted.

The experiments involve speaker independent recognition of 26 letters of the English alphabet. The utterances from the OGI ISOLET database produced by 150 talkers (75 females and 75 males) are used. Each talker utters each of the letters twice. Among them, speech tokens from 120 talkers are used for training and the remaining tokens from the other 30 talkers for testing. To imitate the effect of a telephone bandwidth, the speech data originally sampled at 16 KHz are lowpass-filtered at 3.3KHz and down-sampled to 8 KHz. The feature vectors used in this study consist of 12 bandpass-liftered LPC-derived cepstral coefficients with a 30ms frame length and a 10ms frame shift. In recognition, the decision rule assigns an unknown letter according to the highest forward-backward probability.

Three cases are considered. In the first experiment, a 256-vector codebook is generated from the training tokens of 120 talkers by using the LBG algorithm with a Euclidean distortion measure and used in all experiments. For each letter, at first, a left-to-right 5-state discrete HMM (DHMM) with arbitrary state skipping is trained with the traditional Baum-Welch algorithm. Then, by reversing all the training data sequences, a backward DHMM with the same structure as the forward one is also trained for each letter. In the second experiment, 5-state ergodic continuous density HMMs (EG-CDHMM), one forward, one backward, are trained for each letter. The state observation density is assumed to be a single Gaussian density with diagonal covariance matrix. Finally in the third experiment, 5-state left-to-right CDHMMs (LR-CDHMM), one forward and one backward are adopted for each letter. The corresponding close and open test recognition rates are summarized in Table 1.

Table 1 clearly shows that by using the backward HMMs, reasonable recognition rates are also obtained in comparison with their forward counterparts. This fact shows that by assuming the reversed hidden state sequence to follow a Markov chain, the resultant backward HMMs are also useful for speech recognition, albeit the theoretical fact that a Markov chain which is usually assumed in forward HMMs, looked at in reverse order, in general, is not a Markov chain. This observation may also suggest that the speech model which captures contextual information need not be highly accurate for it to be useful. What is critical for speech modeling may lie in the way to account for local temporal interactions of the underlying state sequence as well as the feature vectors themselves.

With the above observation, one wants to know if there exists the potential of using some techniques to combine the two recognizers with corresponding forward and backward HMMs into one whose accuracy is greater than either individual recognizer can obtain. The scenario is like this: For each speech unit one has two models (λ for forward HMM and $\tilde{\lambda}$ for backward HMM). Thus, given an unknown utterance $\mathbf{X}$, corresponding to each speech unit, two opinions are encoded as two probability distributions ($p(\mathbf{X}|\lambda)$ and $p(\mathbf{X}|\tilde{\lambda})$) to represent the degrees of belief that the unknown utterance belongs to this speech unit. The problem now becomes how to aggregate the above two opinions to enhance the individual's decision based on a single evidence of belief. One possibility to combine the probability distributions is to use so called *linear opinion pool*. The discriminant function of the recognizer can be defined as

$$T(\mathbf{X}|\lambda,\tilde{\lambda}) = \alpha \cdot p(\mathbf{X}|\lambda) + (1-\alpha)\cdot p(\mathbf{X}|\tilde{\lambda}) , \qquad (3)$$

where α is a non-negative weight. With this discriminant function, a series of recognition experiments are conducted to examine the effects of different weighting values of α with the forward and backward DHMMs trained in the previous experiments. The same α is used for all speech units. The related recognition rates are tabulated in Table 2. As for how to determine the optimal values of the weighting coefficient α, one may use the similar technique of the so called *deleted interpolation*, which has been used popularly and successfully in HMM smoothing to solve the insufficient training data problem.

Table 2 clearly shows that with an appropriate weighting coefficient, the combined recognizer has a recognition accuracy greater than that of either one-sided HMM. The method benefits from the consideration of the bi-directional contextual dependence information which are critical for speech modeling. So far as this is concerned, one wants to know if there exists an integrated formulation of this kind of information. Because MRF models provide an excellent vehicle for blending information about local temporal interaction into a global framework [1, 2], in the following section, such kind of models will be explored. However, note that according to the existing literature on MRFs, it is still an open problem to estimate the parameters of hidden MRFs and to get the global maximum *a posteriori* (MAP) estimate of the most likely hidden state sequence corresponding to an observed feature vector sequence. The difficulties stem from the conditional distribution structure of MRFs which makes the computational demand for the Monte Carlo simulation of MRFs very substantial. As a consequence, many approximate methods related to parameter estimation and state labeling have been developed. Also noted is the fact that in this paper, an MRF is not viewed as a definitive representation of a speech utterance, but rather, as a locally dependent MRF serves only to provide a way of formalizing the motivation of utilizing the contextual dependence in a hidden state sequence.

4. MRF BASED ACOUSTIC MODELING

4.1. CVQ Based Acoustic Modeling

With the motivation of utilizing the bi-directional contextual dependence in a hidden state sequence, we have previously proposed a contextual VQ (CVQ) method based on the MRF theory to model the speech feature vector space [3]. Its superiority over other alternative methods is confirmed by a series of comparative experiments in a speaker independent isolated word recognition task by using different VQ schemes as the front-end of DHMM. In

Table 2: The effects of combined recognizer with the forward and backward HMMs

Weight α	1.0	0.9	0.8	0.7	0.6	0.5	0.4	0.3	0.2	0.1	0.0
Close-test	85.0	85.4	86.1	86.6	86.8	86.9	84.6	86.0	85.2	84.1	82.5
Open-test	73.9	75.5	75.8	76.0	75.9	75.8	75.2	74.2	72.8	71.3	69.9

this section, a preliminary study is conducted to extend CVQ method for acoustic modeling of the basic speech units in speech recognition.

Let $\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_T$ be an observed speech feature vector sequence corresponding to a basic speech unit. Each speech unit is modeled by a stochastic machine (model) $\lambda = \{\mathcal{G}, \mathcal{P}, \mathcal{F}\}$ with the associated CVQ codebook $\mathcal{G} = \{G_k, k = 1, 2, \cdots, K\}$. $\mathcal{P} = \{\omega_i, p_{ij}, q_{ik}\}$ represents the stationary prior distribution $Pr(z_t)$ and the directional transition probabilities $Pr(z_{t\pm1}|z_t)$ where $\omega_i = Pr(z_t = G_i)$, $p_{ij} = Pr(z_{t+1} = G_j|z_t = G_i)$ and $q_{ik} = Pr(z_{t-1} = G_k|z_t = G_i)$. $\mathcal{F} = \{f_i(\mathbf{x};\theta_i)\}$ represents the set of PDFs corresponding to each codeword G_i with the associated parameters denoted as θ_i where $f_i(\mathbf{x}_t;\theta_i) = p(\mathbf{x}_t|z_t = G_i)$. Here, for simplicity, each $f_i(\mathbf{x};\theta_i)$ is assumed to have a normal PDF $\mathcal{N}(x|\mu_i, \Sigma_i)$ with μ_i being the D-dimensional mean vector and Σ_i being the $D \times D$ covariance matrix.

A CVQ labeling method is as follows: given an observation sequence $\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_T$, assign each $\mathbf{x}_t$ to codeword (or state) G_i if

$$G_i = \underset{z_t}{\operatorname{argmax}} Pr(z_t) \cdot p(\mathbf{x}_t|z_t) \cdot \left\{ \sum_{z_{t-1}} Pr(z_{t-1}|z_t) \cdot p(\mathbf{x}_{t-1}|z_{t-1}) \right\} \cdot \left\{ \sum_{z_{t+1}} Pr(z_{t+1}|z_t) \cdot p(\mathbf{x}_{t+1}|z_{t+1}) \right\}, \quad (4)$$

where the last two bracketed terms represent the contribution of contextual information. For labeling of boundary feature vectors, say, $\mathbf{x}_1$ and $\mathbf{x}_T$, only its right or left contextual information respectively are used. Although alternative methods exist, a so called decision-directed (DD) method was adopted to estimate the CVQ model parameters and readers are referred to [3] for details of the CVQ training algorithm.

By using the CVQ method, one can generate a CVQ model for each speech unit. Let's consider a collection of M such models, $\Lambda = (\lambda_1, \lambda_2, \cdots, \lambda_M)$, where λ_m denotes the set of parameters of the m-th model, one for each speech unit. Given an unknown utterance $\mathbf{X} = (\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_T)$, the corresponding missing state sequence $\mathbf{Z}^{(i)} = (z_1^{(i)}, z_2^{(i)}, \cdots, z_T^{(i)})$ is identified with the CVQ labeling method using each CVQ model in turn. Then, with the discriminant function defined for class C_i as the accumulated likelihood function

$$g_i(\mathbf{X};\lambda_i) = \prod_{t=1}^{T} f(\mathbf{x}_t|z_t^{(i)}), \quad (5)$$

speech recognition can be done by making the decision for each input $\mathbf{X}$ by choosing the largest of the discriminants evaluated on $\mathbf{X}$ which is often generically stated as

$$C(\mathbf{X}) = C_i, \quad \text{for } g_i(\mathbf{X};\lambda_i) = \max_j g_j(\mathbf{X};\lambda_j), \quad (6)$$

where $C(\cdot)$ denotes a classification operation.

4.2. ICM Based Acoustic Modeling

In MRF-based acoustic modeling for speech recognition, to find the "optimal" state sequence $\mathbf{Z}$ associated with the given observation sequence $\mathbf{X}$, apart from the above CVQ labeling algorithm, a more general labeling algorithm called ICM (iterated conditional modes) algorithm can be adopted [2]. The ICM algorithm is a deterministic iterative algorithm which enables one to find a local maximum of the posterior distribution $Pr(\mathbf{Z}|\mathbf{X},\lambda)$ and thus to get an approximate MAP (maximum *a posteriori*) estimate of the hidden state sequence. More specifically, the ICM algorithm is an iterative algorithm to decompose the global optimization problem

$$\mathbf{Z}^* = \underset{\mathbf{Z}}{\operatorname{argmax}} Pr(\mathbf{Z}|\mathbf{X},\lambda) \quad (7)$$

into the following local optimization problems

$$z_t^{(n+1)} = \underset{z_t}{\operatorname{argmax}} p(\mathbf{x}_t|z_t) Pr(z_t|z_{\eta_t}^{(n)}), \quad (8)$$

where η_t is the neighborhood of the site t and $z_{\eta_t}^{(n)}$ is a provisional estimate of z_{η_t} at the n-th updating. When applied to each frame (site) in turn, this procedure defines a single cycle of an iterative algorithm for estimating $\mathbf{Z}^*$. As an initial estimate $\mathbf{Z}^{(0)}$, one can normally adopt the conventional maximum likelihood classifier which merely choose $z_t^{(0)}$ to maximize $p(x_t|z_t)$ at each t separately. One then applies the algorithm for a fixed number of cycles or until convergence (no changes of labels in two consecutive cycles) to produce the final labeling $\mathbf{Z}^*$. In practice, convergence seems extremely rapid, with few if any changes occurring after several cycles (specifically here 4 cycles).

The above updating procedure represents one extreme. In practice, other more efficient updating schemes can also be adopted, but convergence can no longer be guaranteed and small oscillations may occur. Another extreme is to use synchronous updating whereby each new estimate of z_t is calculated in turn according to equation (8), but the current estimate of $\mathbf{Z}$ is not updated until each cycle is completed. The partially synchronous scheme, in which *coding sets* of sites [1] (specifically here two sets: odd-number frames and even-number frames) are simultaneously updated, provides a useful compromise. This procedure is also very suitable for the parallel processing implementation, provided a single processor is dedicated to each frame, and it is adopted in this study.

To use above ICM algorithm for acoustic modeling, similar formulation can be adopted as in CVQ case. The only difference is that in ICM formulation with first-order MRF considered, the local conditional distribution $Pr(z_t|z_{t-1}, z_{t+1})$, instead of bi-directional transition probabilities as in CVQ case, is used in hidden state sequence labeling. Although specifically designed Gibbs distribution should be used in MRF modeling, for simplicity, in this study we assume $Pr(z_t = G_i|z_{t-1} = G_j, z_{t+1} = G_k)$ is stationary and represented directly by p_{ijk} as model parameters. As for the ICM model parameter estimation, the following decision-directed training method is adopted.

Step 1. Given n training feature vectors, obtain an initial estimate of $\{p_{ijk}\}$, $\{\mu_i\}$ and $\{\Sigma_i\}$ with conventional methods, also obtain an initial estimate of hidden state labels.

Step 2. Based on the current estimate of the model parameters and the state labeling, carry out a single cycle of the ICM labeling algorithm on the training data set to obtain: a new state labeling, n_{ijk} = *count of* $\{z_{t-1} = G_j, z_t = G_i, z_{t+1} = G_k\}$ and n_i = *the number of training vectors assigned to state* G_i.

Step 3. Update the parameters as follows:

$$\hat{p}_{ijk} = n_{ijk} / \sum_i n_{ijk} \quad (9)$$

$$\hat{\mu}_i = \frac{1}{n_i} \sum_{j=1}^{n_i} \mathbf{x}_j^{(i)} \quad (10)$$

$$\hat{\Sigma}_i = \frac{1}{n_i} \sum_{j=1}^{n_i} (\mathbf{x}_j^{(i)} - \hat{\mu}_i)(\mathbf{x}_j^{(i)} - \hat{\mu}_i)^t, \quad (11)$$

where $\mathbf{x}_j^{(i)}$ denotes the jth training vector assigned to state G_i.

Step 4. Repeat steps 2 and 3 for a fixed number of cycles or until convergence (e.g. no change of labeling results in two consecutive cycles or the variation of the total likelihood is less than a predefined threshold).

Note that the restriction that at least $D + 1$ observations belong to each G_i is needed to avoid the degenerate case of infinite likelihood. Due to the insufficient training data problem, smoothing of p_{ijk} is sometimes necessary in practice. The simple threshold method is adopted in this study: if p_{ijk} are less than a predefined threshold (e.g. 10^{-6}), they are set to the threshold value and renormalized to comply with the stochastic constraints.

4.3. Experimental Result

To examine the effects of the hidden MRF-based method for speech recognition, a series of comparative experiments are conducted. The recognition task is the same as the one in previous experiments in section 3. For each letter, the LBG codebook with 5 codewords is first generated. Then, 5-state CVQ and ICM models are trained. The related experimental results are summarized in Table 3. For comparison purpose, the recognition rates with 5-state ergodic CDHMMs are also listed (simply labeled as HMM). The fact that the performance with MRF-based methods is slightly better than the other methods confirms that the proposed methods which take into account the bi-directional contextual dependence information are viable for acoustic modeling of the basic speech unit in speech recognition. Further research along this line of thought are definitely needed to make this kind of methods applicable to a real-life speech recognition system. One possibility is to adopt the parametric distribution such as the Gibbs distribution which is expected to use less parameters to model the hidden sate sequence. Another possibility is to use MRF to directly model the local correlation of the speech feature vectors.

5. SUMMARY AND DISCUSSIONS

With the motivation of utilizing bi-directional contextual dependence in acoustic modeling, in this paper, a bi-directional hidden Markov modeling approach for speech recognition is studied and the importance of bi-directional contextual dependence information for speech recognition is identified by a series of comparative experiments. Furthermore, hidden MRF-based acoustic modeling techniques using a previously proposed CVQ method and ICM algorithm which is very suitable for parallel processing implementation are also attempted. Their viability is confirmed by a series of preliminary experiments. In view of the fact of the bi-directional contextual dependent nature in the speech production process as stated in the introduction section, the motivation and techniques presented in this paper can also be applied to statistical and stochastic language modeling at different level to take advantage of the bi-directional constraint which apparently exists in any natural language. Despite the intuitive appeal of the motivation to use MRF to model contextual dependence in the underlying speech production process, its potential usability in speech modeling will depend critically on the solution of some fundamental problems which include the selection of the appropriate MRFs for speech modeling, the efficient parameter estimation approach, and the computationally effective method for the MAP identification of the hidden state sequence. What is important at this point is not the techniques that have been presented here, but rather an understanding of the problem, from which one can hope that more sophisticated techniques will evolve.

Table 3: Performance comparison (% correct) of MRF-based approach and other methods

Methods	LBG	HMM	CVQ	ICM
close-test	62.0	66.9	68.1	68.2
open-test	60.1	65.1	65.6	65.8

ACKNOWLEDGEMENT

The authors would like to thank Mr. Q.-Y. Feng and Mr. Z.-Q. Ma for helping to conduct the related experiments. The first author would also like to thank Drs. Y. Yamazaki & Y. Sagisaka of ATR-ITL for their support of his preparing this manuscript.

REFERENCES

[1] J. Besag. "Spatial interactions and the statistical analysis of lattice systems". *J. Roy. Stat. Soc. B*, Vol. 36, pp.192-236, 1974.

[2] J. Besag. "On the statistical analysis of dirty pictures". *J. Roy. Stat. Soc. B*, Vol. 48, pp.259-302, 1986.

[3] Q. Huo & C. Chan. "Contextual vector quantization for speech recognition with discrete hidden Markov model". *Pattern Recognition*, Vol. 28, pp.513-517, 1995.

[4] H. Lucke. *Improved acoustic modeling for speech recognition using 2D Markov random fields.* Procs. ICASSP, 1995, pp.540-543.

[5] H. Noda & M. N. Shirazi. *A MRF-based parallel processing algorithm for speech recognition using linear predictive HMM.* Procs. ICASSP, 1994, pp.I-597-600.

[6] Y.-X. Zhao, L. E. Atlas & X.-H. Zhuang, "Applications of the Gibbs distribution to hidden Markov modeling in speaker independent isolated word recognition". *IEEE Trans. on Signal Processing*, Vol. 39, pp.1291-1299, 1991.