A NEW COMBINED MODEL OF STATICS-DYNAMICS OF SPEECH

Zhijian Ou, *Zuoying Wang*

Department of Electronic Engineering, Tsinghua Univ., Beijing 100084, P.R.China
Email: ozj@thsp.ee.tsinghua.edu.cn

ABSTRACT

Linear prediction (LP) HMM does not make the independent and identical distribution (IID) assumption in traditional HMM; however it often produces unsatisfactory results. In this paper, a new combined model of statics-dynamics of speech is proposed, based on a new analysis of both HMM's modeling strengths and weaknesses. The new model works with LPHMM as the dynamic part and traditional IID-based HMM as the static part; in addition, easy implementation and low cost are preserved. A new effective re-estimation solution is suggested for parameter tying to achieve better discrimination. Our experiments on speaker-independent continuous speech recognition demonstrated that the combined model achieved 7.5% error rate reduction from traditional HMM.

1. INTRODUCTION

Further improvement of HMM-based speech recognition system has been of great interest and challenge. Traditional HMM assumes the observations within a state are independently and identically distributed (IID); therefore neglects the dynamical spectral information inherent in speech, which is known to be important for accurate acoustic modeling. Augmenting original features with high-order differentials as dynamical information is in common use for this reason. Furthermore, various alternative models are proposed [1] to incorporate dynamics of speech into traditional HMM. However, they suffer from high computation cost and therefore practically have to rely on sub-optimal multi-pass rescoring, which limits their performance on large vocabulary continuous speech recognition (LVCSR).

Remarkably, the approach [2-6] that directly conditions current output on nearby observations with linear prediction (LP) is more attractive than other modeling assumptions, since it is less expensive. In LPHMM, Viterbi decoding and alignment, even the forward-backward algorithm are still applicable with least modification by just low-costly substituting output probability $p(o_t)$ with a correlated one $p(o_t \mid o_{t-1}, \cdots)$. Early works appeared in [2] where no experimental results were reported and [3] where it produced poor results. In [4], it was found "surprisingly" that LPHMM was beneficial for simple cepstral features but not for features augmented with differentials, and "paradoxically" that LPHMM produced poor recognition rate although the likelihood obtained was much higher than traditional HMM. Our experience with LPHMM for LVCSR also indicated that LPHMM alone is insufficient to describe speech trajectory and often produces undesirable results, in contrast to its theoretic elegance. When combined with discriminant output distributions, LPHMM could reduce the error rate, which was limited to E-set recognition [5]. A marginal dropping of word error rate from 11.8% to 11.4% was reported in [6].

The unsatisfactory and inconsistent performance of LPHMM in practice has not been well understood in the literature. In this paper, a new analysis of LPHMM is provided together with experimental results. The correlated output probability $p(o_t \mid o_{t-1}, \cdots)$ *only* reflects "dynamics of speech", modeling the dynamic variation of each output around some function of nearby observations. On the other hand, IID-based HMM *only* characterizes the "statics of speech", modeling the static location of each output (with the state mean vector) in the feature space. Therefore a natural approach to more accurate acoustic modeling is to integrate these two complementary sources of information together in a combined model.

This paper is organized as follows. In section 2, LPHMM is described; it is further analyzed together with IID-based HMM in section 3, which leads us to propose a new combined model. Experimental results are provided in section 4. Finally the conclusions are made in section 5.

2. LINEAR PREDICTION HMM

Generally suppose the D-dimension observation o_t within a state s is described as

$$o_t = \sum_{i=1}^{m} \beta_i^s o_{t+l_i} + \mu_s + \nu_t \,, \tag{1}$$

where l_i is the "offset" associated with the i^{th} predictor, $\beta_i^s \in R^{D \times D}$ is the i^{th} prediction matrix, $\mu_s \in R^D$ explicitly accounts for a non-zero mean of the observations, and $\nu_t \sim \mathrm{N}(0, \Sigma_s)$ is zero mean full covariance gaussian noise which is un-correlated between frames. The output probability density function (pdf) for the state s and observation o_t then becomes correlated and is given by

$$\begin{aligned} \tilde{b}_s(o_t) &\stackrel{\Delta}{=} p(o_t \mid o_{t+l_i}, i = 1, \cdots, m, s) \\ &= \frac{1}{(2\pi)^{D/2} \left|\Sigma_s\right|^{1/2}} \exp\left\{ -\frac{1}{2} \left(w_t^s - \mu_s\right)^T \Sigma_s^{-1} \left(w_t^s - \mu_s\right) \right\}, \end{aligned} \tag{2}$$

where $w_t^s = o_t - \sum_{i=1}^{m} \beta_i^s o_{t+l_i}$. The probability evaluation in LPHMM is the same as in traditional HMM except that the standard output pdf $b_s(o_t)$, now as gaussian with full covariance $\mathrm{N}(m_s, \Lambda_s)$, is replaced by a correlated one $\tilde{b}_s(o_t)$. The re-estimation of model parameters specific to LPHMM in Viterbi training is derived as follows.

2.1. Parameter Re-estimation in LP-HMM

After Viterbi alignment of the training data $O = o_1 o_2 \cdots o_T$ against the labels $s_1 \cdots s_N$, we obtain the most likely state sequence Q_λ under current model parameters $\lambda = \{\mu_s, \Sigma_s, \beta_i^s, i = 1, \cdots, m\}$ and thus a frame-set Γ_s for each state-label s that specifies which frames are assigned to s after segmentation. The log-likelihood to be maximized is

$$L(\hat{\lambda}) = \log P(O \mid Q_\lambda, \hat{\lambda}) = \sum_s \sum_{t \in \Gamma_s} \log \tilde{b}_s(o_t \mid \hat{\lambda}). \quad (3)$$

A straightforward maximization by differentiating $L(\hat{\lambda})$ with respect to all model parameters gives the following re-estimation formulae.

$$\hat{\mu}_s = \frac{\Phi_0^s - \sum_{i=1}^m \hat{\beta}_i^s \Phi_i^s}{|\Gamma_s|}, \quad (4)$$

$$\hat{\Sigma}_s = \frac{R_{00}^s - \sum_{i=1}^m \left(R_{0i}^s \hat{\beta}_i^{s\,T} + \hat{\beta}_i^s R_{0i}^{s\,T} \right) + \sum_{i=1}^m \sum_{j=1}^m \hat{\beta}_i^s R_{ij}^s \hat{\beta}_j^{s\,T}}{|\Gamma_s|} - \hat{\mu}_s \hat{\mu}_s^T \quad (5)$$

$$\left(\hat{\beta}_1^s, \cdots, \hat{\beta}_m^s \right) = \left(\hat{B}_1^s, \cdots, \hat{B}_m^s \right) \begin{pmatrix} R_{11}^s & \cdots & R_{1m}^s \\ \vdots & \ddots & \vdots \\ R_{m1}^s & \cdots & R_{mm}^s \end{pmatrix}^{-1}, \quad (6)$$

where $\Phi_i^s = \sum_{t \in \Gamma_s} o_{t+l_i}$, $R_{ij}^s = \sum_{t \in \Gamma_s} o_{t+l_i} o_{t+l_j}^T$, $\hat{B}_i^s = \sum_{t \in \Gamma_s} (o_t - \hat{\mu}_s) o_{t+l_i}^T$, $0 \le i, j \le m$ are gathered statistics specific to state s.

3. A COMBINED MODEL

3.1. Analysis of LPHMM and IID-based HMM

Another different approach to maximize (3) helps us to gain some insight into the property of LPHMM. For convenience, let $x_t = \left(o_t^T, o_{t+l_1}^T, \cdots, o_{t+l_m}^T \right)^T$, $\theta_s = \left(I, -\beta_1^s, \cdots, -\beta_m^s \right)$, then $w_t^s = \theta_s x_t$, and $L(\hat{\lambda})$ can be re-written as

$$L(\hat{\lambda}) = -\sum_s \frac{|\Gamma_s|}{2} \left\{ \log \left| \hat{\Sigma}_s \right| + trace \left[\hat{\Sigma}_s^{-1} \left(\hat{\theta}_s C_x^s \hat{\theta}_s^T \right) \right] + \left(\hat{\theta}_s \eta_x^s - \hat{\mu}_s \right) \hat{\Sigma}_s^{-1} \left(\hat{\theta}_s \eta_x^s - \hat{\mu}_s \right)^T + D \log(2\pi) \right\} \quad (7)$$

where η_x^s, C_x^s are the sample mean and covariance respectively calculated on the data set $\{x_t \mid t \in \Gamma_s\}$, i.e. the extended-frames assigned to s. *First*, we can obtain the mean and covariance estimates, $\hat{\mu}_s = \hat{\theta}_s \eta_x^s$, $\hat{\Sigma}_s = \hat{\theta}_s C_x^s \hat{\theta}_s^T$, that maximize the likelihood $L(\hat{\lambda})$ in terms of fixed prediction matrices $\hat{\theta}_s$. *Then* substituting $\hat{\mu}_s$ and $\hat{\Sigma}_s$ into (7), we have the log-likelihood in terms of only $\hat{\theta}_s$,

$$L(\{\hat{\theta}_s\}) = -\sum_s \frac{|\Gamma_s|}{2} \left\{ \log \left| \hat{\theta}_s C_x^s \hat{\theta}_s^T \right| + D \log(2\pi) + D \right\}. \quad (8)$$

To estimate prediction matrices, the likelihood function (8) is maximized, or equivalently $\left| \hat{\theta}_s C_x^s \hat{\theta}_s^T \right|$ is minimized, which is the determinant of the sample covariance of $o_t - \sum_{i=1}^m \beta_i^s o_{t+l_i}$. Since the determinant of the sample covariance of a random variable provides a good measure of how compact it is distributed, to minimize $\left| \hat{\theta}_s C_x^s \hat{\theta}_s^T \right|$ is to find such β_i^s's that o_t is most compactly distributed conditional on the context (or say, around the value of $\sum_{i=1}^m \beta_i^s o_{t+l_i}$). In this way, the *dynamics* of outputs of state s is well captured in LPHMM embodied by the correlated output pdf $\tilde{b}_s(o_t)$. On the other hand, IID-based HMM is still effective in practical speech recognition, maybe due to its good ability at modeling the *statics* of speech. All the observations in each state are well statically (unconditionally) distributed in a cluster represented by the mean of the standard output pdf $b_s(o_t)$, regardless of any nearby observations.

The weak points are that, to decide which state the feature o_t most probably comes from, the matching score computed by $\tilde{b}_s(o_t)$ alone is insufficient, if the matching score by $b_s(o_t)$ is not taken into account, and vice versa. Regarding o_t as one-dimensional, Fig. 1(2) illustrates the former (later) case with two states. Each ellipse is the contour line of $p(o_t, o_{t-1} \mid s)$, characterizing the output features of the state s=1,2. The gaussian pdf curves along the o_t axis and the sloping line l_1 (perpendicular to $o_t = \beta^1 o_{t-1}$) respectively represent $\{b_1(o_t), b_2(o_t)\}$, and $\{\tilde{b}_1(o_t), \tilde{b}_2(o_t)\}$ that are put together along l_1 for clear view. The overlapping area of two gaussian pdf curves gives the classification error. To make statistical decisions, using $b_s(o_t)$ and $\tilde{b}_s(o_t)$ yields Err_s and Err_d respectively. When $Err_s > Err_d$, the distribution of *statics* represented by $b_s(o_t)$ is more discriminative than the distribution of *dynamics* represented by $\tilde{b}_s(o_t)$, and vice versa.

3.2. Combine LPHMM and IID-based HMM

It is beneficial to utilize the complementary modeling powers on statics and dynamics of speech of these two kinds of HMM's to yield a combined model, since quasi-stationary statics and transitional dynamics are actually mixed in any segment of speech. The new "combined output pdf" is defined as

$$\tilde{\tilde{b}}_s(o_t) = b_s(o_t)^{1-\alpha} \cdot \tilde{b}_s(o_t)^{\alpha}, \quad (9)$$

where α is the combination weight. When α=0, 1, the combined model (CM) becomes the traditional HMM and LPHMM respectively.

The CM inherits directly from LPHMM the desirable property that just by replacing $b_s(o_t)$ with $\tilde{\tilde{b}}_s(o_t)$, one can apply to CM as well the decoding and iterative training techniques already developed with low computation cost. In addition, once statistics are obtained, model parameters $\{m_s, \Lambda_s\}$ and $\{\mu_s, \Sigma_s, \beta_i^s, i = 1, \cdots, m\}$ are actually re-estimated separately. Thus the implementation of CM requires least changes in the trainer's and decoder's design.

3.3. Parameter tying for better discrimination

When applying maximum likelihood estimation (MLE) to

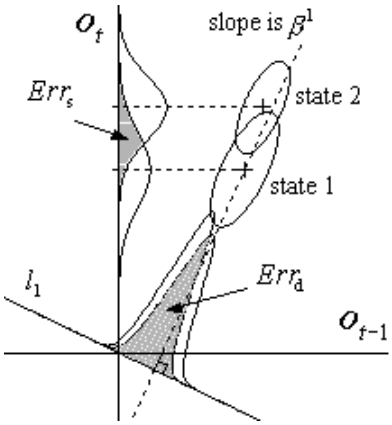

Fig.1 An illustration of how LPHMM fails to discriminate between two states, where $Err_s < Err_d$.

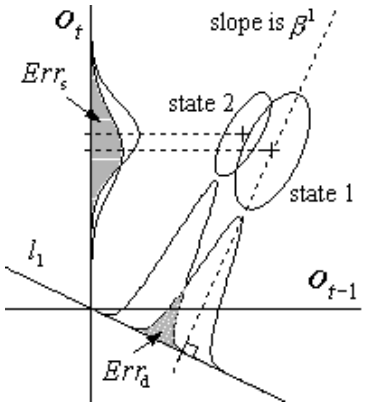

Fig.2 An illustration of how IID-based HMM fails to discriminate between two states, where $Err_d < Err_s$.

prediction matrices, the predictors are only trained on positive instances, receiving features in only a small region of the input feature space. When any other features outside are shown, the predictors take it for granted to compute an output as usual, maybe overlapping with the outputs of others, which makes the states fairly confusable. If prediction matrices are tied, this shortcoming can be somehow alleviated, since estimates of parameters are obtained not solely on the features from each state. So it is expected that the performance of LPHMM with tied prediction matrices may be better than the one without tying, which is verified by following experiments. There was similar finding in [7], where constrained MLE with parameter sharing across classes leads to better discrimination.

However, the establishment of (6) requires that the states use state-specific prediction matrices. If prediction matrices are tied across a cluster of states, then (6) holds plus additional common covariance constraint among these states (similar conclusion in [4]) otherwise no analytical solution exists, since

$$\frac{\partial L(\hat{\lambda})}{\partial(\hat{\beta}_1^s,\cdots,\hat{\beta}_m^s)^T}=\begin{pmatrix}\hat{\Sigma}_s^{-1}\sum_{i=1}^m\hat{\beta}_i^s R_{i1}^s\\ \vdots\\ \hat{\Sigma}_s^{-1}\sum_{i=1}^m\hat{\beta}_i^s R_{im}^s\end{pmatrix}-\begin{pmatrix}\hat{\Sigma}_s^{-1}\hat{B}_1^s\\ \vdots\\ \hat{\Sigma}_s^{-1}\hat{B}_m^s\end{pmatrix}.$$

Simply pooling the statistics [5] belonging to a state-cluster required in (6) for computing tied $(\hat{\beta}_1^{\,c},\cdots,\hat{\beta}_m^{\,c})$ as defining

$$R_{ij}^c=\sum_{s\in cluster}\sum_{t\in\Gamma_s}o_{t+l_i}o_{t+l_j}^T, \tag{10}$$

$$B_i^c=\sum_{s\in cluster}\sum_{t\in\Gamma_s}(o_t-\mu_s)o_{t+l_i}^T, \tag{11}$$

ignores the effect of covariances. Since correlation between components of the feature vector is mostly modeled by the full covariances of gaussians (i.e. Λ_s,Σ_s), diagonal prediction matrices are preferred in practice. In this case, for tied diagonal prediction matrices, we suggest extending (6) with

$$R_{ij}^c=\sum_{s\in cluster}\left[diag(\hat{\Sigma}_s)\right]^{-1}\sum_{t\in\Gamma_s}o_{t+l_i}o_{t+l_j}^T, \tag{12}$$

$$B_i^c=\sum_{s\in cluster}\left[diag(\hat{\Sigma}_s)\right]^{-1}\sum_{t\in\Gamma_s}(o_t-\mu_s)o_{t+l_i}^T, \tag{13}$$

thereafter still applying (4)(5). If diagonal gaussian mixtures are used, the above solution moreover stands analytically, and increases likelihood by each step, since

$$L(\{\beta_i^{\,s},\mu_s,\Sigma_s\})\le L(\{\hat{\beta}_i^{\,s},\mu_s,\Sigma_s\})\le L(\{\hat{\beta}_i^{\,s},\hat{\mu}_s,\hat{\Sigma}_s\}).$$

3.4. Discussions

Examples of the above log-linear combination for fusing several sources of information can be found in the use of "language model factor" when multiplying the acoustic model score with the language model score, and the use of "codebook exponent" when combing multiple codebooks on different parameter sets. The bigram-constrained (BC) HMM [9] falls as a special case of the above CM, when forcing $\alpha=0.5$, $m=1$, $l_1=-1$ and all the prediction matrices are tied to be a global state-independent one. Similar to our CM, a *priori-posteriori* combination of probability distributions (PD) appeared in [10]. It was found and heuristically explained that the a *posteriori* PD, analogous to $\tilde{b}_s(o_t)$ but based on extended logarithmic pool (ELP), is insufficient alone for recognition. If so, the previous complicated step to obtain the a *posteriori* PD based on ELP seems redundant. Experiments only on discrete HMM were reported. The proposed CM is more flexible, efficient and theoretically sound.

4. EXPERIMENTAL RESULTS

To assess the effectiveness of CM and to demonstrate the points made previously, experiments were carried on a speaker-independent LVCSR task using the male speech database for "National 863 Assessment". Utterances from 76 speakers were used as training data and those from the other 7 speakers formed the test data, with each speaker about 600 sentences. In the front-end the speech was parameterized into 14 MFCCs along with normalized log-energy, and their first and second order differentials. The system employed the semi-syllable units, including 100 Initial units each with 2 states, 164 Final units each with 4 states, plus one single-state silence model. Recognition experiments below gave acoustic recognition results in the form of tone-syllable. The (syllable) error rate is defined as the percentage of the sum of numbers of syllables decoded as deletions, insertions and substitutions. The baseline system gave 26.30% error rate. When combined with language model, the recognition system yielded 4% character error rate. All the following experiments used diagonal prediction matrices.

The average error rates of LPHMM and the CM with different configurations of the predictors are listed in Table 1. The "Offsets" column implicitly specifies m. Here prediction

Baseline: **26.30**%

Offsets	LPHMM	CM
–1	43.77	25.75
–2	32.67	25.50
–3	27.26	25.18
–4	25.24	**24.32**
+4	25.82	24.92
–3, –1	44.46	25.15
–4, –2	32.74	25.02
–4, +4	25.57	24.76

Table 1: Average error rate of LPHMM and CM (α=0.3) as a function of predictor offsets with prediction matrices tied across all states

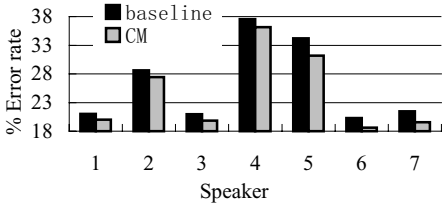

Fig.3: The error rate of the baseline and CM on test speakers with a global –4 predictor, α=0.3

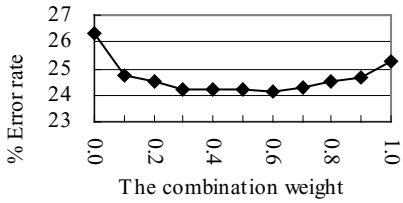

Fig.4 The error rate of CM as a function of α with a global –4 predictor

Different search-space pruning methods	Nearly full search constrained by $\tau^s_{\max}$	*MBest* search , *M* =			State boundaries constraint ±1
		1	2	3	
Computation cost ≈	15 (averaged $\tau^s_{\max}$)	2	3	4	3
poly-fitting with order 1	25.65	26.09	25.77	25.68	26.02
poly-fitting with order 2	25.42	25.95	25.63	25.49	25.71

Table 2: The error rate of polynomial-fitting trajectory model [11]. The second row shows how many times the cost of the baseline the models have. $\tau^s_{\max}$ is the maximum state duration.

matrices were tied across all states, and α=0.3. These results demonstrate unsatisfactory recognition performance by using LPHMM alone, which is hardly better than the baseline. The CM was consistently much better than both the baseline and LPHMM in all cases, which clearly indicates its superiority over the other two models. The complementary modeling of the statics and dynamics of speech indeed improves recognition performance in practice, which agrees with our previous analysis. The best result was obtained using a single diagonal predictor at an offset of –4, and the relative error rate reduction from the baseline was 7.5%. It should be emphasized that the error rates with CM for all test speakers were uniformly cut down as shown in Fig 3.

Next we assessed how the weight α affects the performance of CM. Fig. 4 shows the error rate of CM with a global –4 predictor as a function of α. It can be seen that gains achieved by CM were not sensitive to varying α in some range, which implies that the coupling in CM is stable.

The above all experiments used the re-estimation solution suggested in (12,13). Ignoring the covariance to use (10,11), the CM under a global –4 predictor with α=0.3 yielded the error rate of 25.17% > 24.32%, which shows the advantage of the suggested methods.

Experiment using separate prediction matrices for each state was also taken under a single –4 predictor with α=0.3. Not surprisingly, CM produced the error rate of 26.14% greater than 24.32% achieved with tied prediction matrices as above. This is another supporting evidence that parameter sharing may alleviate the shortcoming of MLE and lead to good discrimination.

Finally comparison was made between CM and the known polynomial-fitting trajectory model as in [11]. The results are summarized in Table 2. It is clear that CM has advantage over the polynomial-fitting trajectory model, not only by greater error rate reduction but also by less computation cost with only about two times the baseline.

5. CONCLUSIONS

Motivated by an analysis of LPHMM, a new combined model of statics-dynamics of speech is proposed, which overcomes the weakness of IID assumption in traditional HMM, with easy implementation and low cost. Experiments on a speaker-independent LVCSR task showed its advantages over both LPHMM and IID-based HMM, with great error rate reduction. Furthermore, different degree of tying of prediction matrices may be beneficial, not just tied globally or separately for each state. The combination weight α can be state-dependent and optimized with MMI training. Further works on the above issues are promising.

REFERENCES

[1] M. Ostendorf et al., "From HMM's to segment models: a unified view of stachastic modeling for speech recognition", IEEE Trans. on SAP, vol.4, no.5, 1996.

[2] C.J. Wellenkens, "Explicit correlation in hidden Markov model for speech recognition", Proc. ICASSP 1987.

[3] P.F. Brown, 'The acoustic modeling problem in automatic speech recognition", IBM Tech. Report, No. RC 12750, 1987.

[4] Kenny, et al., "A linear predictive HMM for vector-valued observation with application to speech recognition", IEEE Trans. on ASSP, vol.38, no.2, 1990.

[5] P.C. Woodland, "Hidden Markov models using vector linear prediction and discriminative distributions", Proc. ICASSP 1992.

[6] Y. Jia, J. Li, "Relax frame independence assumption for standard HMMs by state dependent auto-regressive feature models", Proc. ICASSP 2001.

[7] R.A.Gopinath, "Maximum likelihood modeling with gaussian distributions for classification", Proc. ICASSP 1998.

[8] Y. Normandin, et al., "High-performance connected digit recognition using maximum mutual information estimation", IEEE Trans.on SAP, vol.2, no.2, 1994.

[9] S. Takahashi, et al., "Phoneme HMM's constrained by frame correlations", Proc. ICASSP 1993.

[10] N.S. Kim, et al., "Frame-correlated hidden Markov model based on extended logarithmic pool", IEEE Trans. on SAP, vol.5, no.2, 1997.

[11] L.Deng, "Speech recognition using hidden Markov models with polynomial regression functions as nonstationary states", IEEE Tran. on SAP, vol.2, no.4, 1994.