A COMBINED MODEL OF STATICS-DYNAMICS OF SPEECH OPTIMIZED USING MAXIMUM MUTUAL INFORMATION

Zhijian Ou, *Zuoying Wang*

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

ABSTRACT

The linear prediction (LP) HMM does not make the independent and identical distribution (IID) assumption in the traditional HMM; however it often produces unsatisfactory results. In our previous paper [7], both HMMs' modeling strengths and weaknesses were analyzed and a new combined model of statics-dynamics of speech was proposed. It works with LPHMM as the dynamic part and with the traditional IID-based HMM as the static part; in addition, easy implementation and low cost are preserved. In this paper, an optimal combination using maximum mutual information (MMI) is introduced. Our experiments on speaker-independent continuous speech recognition demonstrated that the combined model achieved better performance than both models.

1. INTRODUCTION

It has been well known that a major limitation to the traditional HMM in speech modeling is that the observations within a state are assumed to be independently and identically distributed (IID), which neglects the useful dynamic spectral information inherent in speech. Various alternative models are proposed [1] to incorporate dynamics of speech into the traditional HMM. However, they suffer from high computation cost and therefore practically have to rely on sub-optimal multi-pass rescoring and pruning, 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. 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 the traditional HMM. 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 [7], we proposed a new combined model of statics-dynamics of speech. 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. Thus it is beneficial to integrate these two complementary sources of information together in a combined model. Preliminary experimental results were very encouraging [7]. In this paper, we introduce an optimal combination, which uses maximum mutual information (MMI) to estimate the combination weight. Recent experimental results using the combined model with and without such optimization technique are reported.

This paper is organized as follows. In section 2, LPHMM is briefly described. We present the combined model in section 3, and introduce the optimal combination using MMI in section 4. Experimental results are provided in section 5. Finally the conclusions are made in section 6.

2. LINEAR PREDICTION HMM

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

$$o_t = \sum_{i=1}^{m} \beta_i^s o_{t+l_i} + \mu_s + v_t , \quad (1)$$

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 $v_t \sim \mathrm{N}(0, \Sigma_s)$ is zero mean full covariance gaussian noise which is un-correlated between frames. For state s, the output probability density function (pdf) of observation o_t then becomes correlated conditional on its context $\{o_{t+l_1}, \cdots, o_{t+l_m}\}$:

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

where $w_t^s = o_t - \sum_{i=1}^{m} \beta_i^s o_{t+l_i}$.

3. THE COMBINED MODEL

3.1. Analysis and motivation

An understanding of how observations o_t's are modeled respectively in IID-based (i.e., traditional) HMM and LPHMM is illustrated in Fig. 1. There o_t is regarded as one-dimensional and $m = 1$, $l_1 = -1$. Each ellipse is the contour line of $p(o_t, o_{t-1} \mid s)$, conceptually characterizing the output features of each state s. Throughout the paper, we define the *99% distributed area* (or, for short, distributed area) of a random variable as the smallest elliptical region where the random variable falls with 99%

probability (e.g., for a 1-dimension gaussian variable $\mathrm{N}(m,\sigma^2)$, it is the interval $[m-2.58\sigma, m+2.58\sigma]$).

The distributed area of o_t assumed in LPHMM is the shaded band perpendicular to line $o_t = \beta_1^{\,s} o_{t-1}$. The classification error is determined by the extent of the overlapping of the distributed areas for different states' outputs, which is in turn closely related to both the positions and the widths of the distributed areas belonging to different states. It is now clear that LPHMM does not necessarily perform better than IID-based HMM, since the higher likelihood in LPHMM only indicates its distributed areas are narrower, not necessarily well separated. More rigorously, examining how parameters of LPHMM are chosen helps us gain further insight into its property. The parameters specific to state s are estimated by maximizing the likelihood of the frames assigned to s. Denote $\Gamma_s = \{t \mid q_t = s\}$, $x_t = \left(o_t^{\,T}, o_{t+l_1}^{\,T}, \cdots, o_{t+l_m}^{\,T}\right)^T$, $\theta_s = \left(I, -\beta_1^{\,s}, \cdots, -\beta_m^{\,s}\right)$, and $C_x^{\,s}$ as the sample covariance calculated on the data set $\{x_t \mid t \in \Gamma_s\}$. The problem was shown in [7] to be reduced to a maximization of a new likelihood function,

$$L(\{\theta_s\}) = -\sum_s \frac{|\Gamma_s|}{2}\left\{\log\left|\theta_s C_x^{\,s} \theta_s^{\,T}\right| + D\log(2\pi) + D\right\}, \tag{3}$$

in terms of only θ_s. Equivalently $\left|\theta_s C_x^{\,s} \theta_s^{\,T}\right|$ 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 compactly it is distributed, to minimize $\left|\theta_s C_x^{\,s} \theta_s^{\,T}\right|$ is to find such $\beta_i^{\,s}$'s that o_t is most compactly distributed conditional on its 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 $g(o_t \mid s)$.

On the other hand, the distributed area of o_t assumed in IID-based HMM is the shaded band parallel to o_{t-1}-axis, no matter what nearby o_{t-1} is. All the observations in each state are well statically (unconditionally) distributed in a cluster represented by the mean of the standard output pdf $f(o_t \mid s)$, i.e., gaussian with full covariance $\mathrm{N}(m_s, \Lambda_s)$, regardless of any nearby observations. The traditional HMM is still effective in practical speech recognition, maybe due to its good ability at modeling the *statics* of speech.

3.2 Formulation

Neither LPHMM nor IID-based HMM alone is sufficient. Combination of them provides a more accurate way to model how o_t's are distributed. Our proposed model [7] is to utilize the complementary modeling powers on statics and dynamics of speech of these two kinds of HMMs to yield a combined model. The new "combined output pdf" is defined as

$$p(o_t \mid s) = f(o_t \mid s)^{1-\alpha} \cdot g(o_t \mid s)^{\alpha}, \tag{4}$$

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

4. COMBINATION OPTIMIZED USING MMI

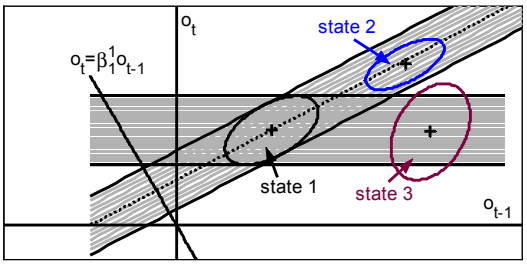

Fig.1: The distributed areas of o_t assumed in LPHMM for state 1 and 2 are heavily overlapped, thus it fails to discriminate between these two states, while IID-based HMM will work well in this case. Consider state 1 and 3 for the counterpart.

Both $f(o_t \mid s)$ and $g(o_t \mid s)$ contribute to the total combined output pdf, yet in different degrees. Previous work [7] used an empirically determined combination weight α. MMI [8] is a good criterion for estimating the combination weight automatically and discriminatively. For lack of discrimination, maximum likelihood (ML) estimation fails here, as it will choose α=1 since the likelihood obtained by LPHMM is generally higher than that by IID-based HMM. More fundamentally, MMI allows us to put (4), though strictly not a pdf, into a statistical maximum *a posterior* (MAP) decision framework.

Theoretically with least probability of error, the MAP decoder is given by $\hat{W} = \arg\max_W P(W \mid O) = \arg\max_W p(O \mid W)P(W)$.

Practically, a parametric model (e.g., HMMs) $p_\lambda(O \mid W)$ is intended as an "estimate" of the true $p(O \mid W)$. And recognition is performed in an approximated MAP sense as,

$$\hat{W} = \arg\max_W p_\lambda(O \mid W)P(W),$$

while assuming a language model is available. Arguably, this is an indirect approach. On the other hand, MMI modeling directly attempts to approximate $P(W \mid O)$, the probability used in the MAP decoding. In this context, the parametric probabilistic model we are interested in is really $P_\lambda(W \mid O)$, expressed as

$$P_\lambda(W \mid O) = \frac{p_\lambda(O \mid W)P(W)}{\sum_W p_\lambda(O \mid W)P(W)}. \tag{5}$$

The *function* $p_\lambda(O \mid W)$ is of interest only to the extent it is used in (5), and in fact not necessarily to be distributions. For N training observations $\{O^{(1)}, \cdots, O^{(N)}\}$ with corresponding transcriptions $\{W^{(1)}, \cdots, W^{(N)}\}$, λ is estimated by maximizing

$$\prod_{n=1}^N P_\lambda\left(W^{(n)} \mid O^{(n)}\right),$$

so as to approximate $P(W \mid O)$ with $P_\lambda(W \mid O)$ as closely as possible. Equivalently the objective function of MMI can be reduced as

$$I_\lambda = \sum_{n=1}^N \left[\log p_\lambda\left(O^{(n)} \mid W^{(n)}\right) - \log p_\lambda\left(O^{(n)}\right)\right], \tag{6}$$

where $p_\lambda\left(O^{(n)}\right) = \sum_W p\left(O^{(n)} \mid W\right)P(W)$.

The summation is efficiently computed by use of word-lattice. Suppose that the recognition result for $O^{(n)}$ is organized as a word-lattice as in Fig. 2, where $w_{l,m}^{(n)}$, $l = 1, \cdots, L, m = 1, \cdots, M$, denotes the m^{th} candidate word at the position l, and M is the

$$\begin{array}{cccc} w_{1,1}^{(n)}, & w_{2,1}^{(n)}, & \cdots & w_{L-1,1}^{(n)}, & w_{L,1}^{(n)} \\ \vdots & \vdots & \cdots & \vdots & \vdots \\ w_{1,m}^{(n)}, & w_{2,m}^{(n)}, & \cdots & w_{L-1,m}^{(n)}, & w_{L,m}^{(n)} \\ \vdots & \vdots & \cdots & \vdots & \vdots \\ w_{1,M}^{(n)}, & w_{2,M}^{(n)}, & \cdots & w_{L-1,M}^{(n)}, & w_{L,M}^{(n)} \end{array}$$

Fig.2: Illustration of a word lattice (Here a word is in fact a Chinese syllable and we use M =100).

number of candidate words at each position. $w_{1,1}^{(n)} w_{2,1}^{(n)} \cdots w_{L,1}^{(n)} = \arg\max_{W} P_\lambda\left(W \mid O^{(n)}\right)$ gives the optimal segmentation of $O^{(n)}$ into $O^{(n)} = O^{(n),1} O^{(n),2} \cdots O^{(n),L}$. Then $p_\lambda\left(O^{(n)}\right)$ is approximated as

$$\begin{aligned} p_\lambda\left(O^{(n)}\right) &= \prod_{l=1}^{L}\left(\sum_{m=1}^{M} p_\lambda\left(O^{(n),l} \mid w_{l,m}^{(n)}\right) P\left(w_{l,m}^{(n)}\right)\right) \\ &= \text{const} \cdot \prod_{l=1}^{L}\left(\sum_{m=1}^{M} p_\lambda\left(O^{(n),l} \mid w_{l,m}^{(n)}\right)\right). \end{aligned} \quad (7)$$

The simplifying assumption is that a uniform unigram is used, which is indeed the case in the *acoustic part* of our system.

Now it is ready to compute the gradient. Suppose $N=1$, drop the variable n for the moment. For multiple training observations, we only need to sum up the gradients computed separately with all observations. For the observation O of length T,

$$\begin{aligned} p_\lambda(O \mid W) &= \prod_{t=1}^{T}\left[f(o_t \mid q_t)^{1-\alpha}\, g(o_t \mid q_t)^{\alpha}\right] \\ &\overset{\Delta}{=} f(O \mid W)^{1-\alpha}\, g(O \mid W)^{\alpha}, \end{aligned}$$

where $f(O \mid W) = \prod_{t=1}^{T} f(o_t \mid q_t)$, $g(O \mid W) = \prod_{t=1}^{T} g(o_t \mid q_t)$ are respectively the likelihood of O with IID-HMM and LPHMM, and $q_1 \cdots q_T$ is the best Viterbi alignment path of the word string W, using the combined pdf. Now we have

$$\frac{\partial I_\lambda}{\partial \alpha} = \log \frac{g(O \mid W)}{f(O \mid W)} - \sum_{l}\sum_{m} \frac{p_\lambda\left(O^l \mid w_{l,m}\right)}{\sum_{m'} p_\lambda\left(O^l \mid w_{l,m'}\right)} \log \frac{g\left(O^l \mid w_{l,m}\right)}{f\left(O^l \mid w_{l,m}\right)}.$$

It is interesting to compute the second derivative. Abbreviating $f_{l,m} = f\left(O^l \mid w_{l,m}\right)$, $g_{l,m} = g\left(O^l \mid w_{l,m}\right)$, then $p_\lambda\left(O^l \mid w_{l,m}\right) = f_{l,m}^{1-\alpha} g_{l,m}^{\alpha}$, and using Cauchy Inequality we have,

$$\begin{aligned} \frac{\partial^2 I_\lambda}{\partial \alpha^2} = -\sum_{l}\Bigg\{ &\left[\left(\sum_m f_{l,m}^{1-\alpha} g_{l,m}^{\alpha}\right)\left(\sum_m f_{l,m}^{1-\alpha} g_{l,m}^{\alpha} \log^2 \frac{g_{l,m}}{f_{l,m}}\right)\right. \\ &\left.\left. - \left(\sum_m f_{l,m}^{1-\alpha} g_{l,m}^{\alpha} \log \frac{g_{l,m}}{f_{l,m}}\right)^2\right] \middle/ \left(\sum_m f_{l,m}^{1-\alpha} g_{l,m}^{\alpha}\right)^2 \Bigg\} \le 0 \end{aligned}$$

Thus I_λ is a concave $\cap$ function of α over $[0,1]$ and will achieve its unique maximum over $[0,1]$. So instead of using gradient ascent method, where the step size is hard to adjust to balance between oscillation and fast convergence, we iteratively bipartition the interval and then select a smaller interval according to the gradient at the midpoint. It in the above case guarantees a convergence precision of 0.01 after only 7 steps.

4.1 Training procedure

There are three parts of parameters $\lambda = \{\lambda_f, \lambda_g, \alpha\}$, where $\lambda_f = \{m_s, \Lambda_s\}$, $\lambda_g = \left\{\mu_s, \Sigma_s, \beta_i^s, i = 1, \cdots, m\right\}$ are respectively parameters specific to IID-HMM as the dynamic part and LPHMM as the static part. Here Viterbi decoding and alignment, are readily applicable with least modification by just low-costly replacing the standard $f(o_t \mid s)$ with the combined $p(o_t \mid s)$. The resulting training procedure of the combined model is as follows.

It is initialized with $\alpha^{(0)} = 0.5$ with interval $\left[L^{(0)} = 0, R^{(0)} = 1\right]$, and $\lambda_f, \lambda_g\left(\beta_i^s = 0\right)$ being set with the parameters of available traditional HMM. Subsequently, each iteration is a two-step process. The *first* step is, leaving α fixed, an alternation of a Viterbi alignment using the combined pdf and an update of λ_f, λ_g, which are actually re-estimated separately once statistics are obtained (See [7] for formula). *Second*, recognition is performed on the training set. The resulting word lattices are used to compute the gradient $I'^{(k)}$ at current $\alpha^{(k)}$. Then $I'^{(k)}$ is used to direct the update as follows.

$$\begin{array}{l} \text{If } I'^{(k)} > 0,\ L^{(k+1)} = \alpha^{(k)}, R^{(k+1)} = R^{(k)}, \text{ and } \alpha^{(k+1)} = \dfrac{L^{(k+1)} + R^{(k+1)}}{2}. \\ \text{If } I'^{(k)} < 0,\ R^{(k+1)} = \alpha^{(k)}, L^{(k+1)} = L^{(k)} \end{array}$$

Iterations are ended when the combination weight converges to a pre-defined precision (e.g., 0.01).

5. EXPERIMENTAL RESULTS

To demonstrate the points made previously, experiments were carried on a speaker-independent Chinese LVCSR task using the male speech database for "China 863 Assessment". Utterances from 76 speakers were used for training and those from the other 7 speakers formed the test data, with about 600 sentences for each speaker.

All Chinese characters are pronounced as one of the total 1254 toned Chinese syllables in CV structure, which are combinations of 100 consonant units and 164 vowel units. A HMM syllable model was used, with 2 states for the consonant part and 4 states for the vowel part.

In the front-end, the speech was parameterized into 14 MFCCs along with the normalized log-energy, and their first and second order differentials. 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 were used. The prediction matrices were state-specific and not tied.

The overall recognition system was composed of two parts. The *acoustic part* decoded the input speech into syllable strings, organized as syllable-lattices, without use of any language model. The subsequent *language part* decoded the syllable strings into Chinese characters. The baseline acoustic model was also an IID-based HMM. Here we focus on the *acoustic part*, and only report the first-candidate syllable error rate.

5.1 Results with fixed $\alpha = 0.5$

As shown in Table 1 and Fig. 3,4, a series of experiments were conducted as a function of different experimental factors: 1) dimension of the features, $D = 15,30(+\Delta),45(+\Delta\Delta)$; 2) configuration of the predictors, $\{l_1, \cdots, l_m\}$. In fact, a different selection of $\{l_1, \cdots, l_m\}$ means a different LPHMM or CM model. For example, $\{-4,+4\}$ stands for such kind of LPHMM (or the dynamic part of CM) that $m = 2$, $l_1 = -4, l_2 = 4$.

LPHMMs had inconsistent performance, heavily depending on the selection of $\{l_1, \cdots, l_m\}$. The performance of LPHMMs with $\{-1\}$, $\{+1\}$ under 15 dimension, $\{-2\}$, $\{-1\}$, $\{+1\}$ and $\{+2\}$ under 30

dim	*model type*	{-5}	{-4}	{-3}	{-2}	{-1}	{+1}	{+2}	{+3}	{+4}	{+5}	{−2,+2}	{−3,+3}	{−4,+4}	{−5,+5}
15	IID-HMM	59.20													
	LPHMM	46.97	45.5	45.28	49.10	63.44	65.15	53.92	45.87	43.51	43.67	86.87	61.43	48.08	44.45
	CM	49.17	47.24	45.01	43.48	43.75	45.13	43.37	44.06	45.04	46.86	41.87	39.67	39.92	41.40
30	IID-HMM	29.59													
	LPHMM	28.21	28.21	29.85	33.90	44.11	43.43	34.11	29.70	28.20	28.11	58.88	35.42	29.81	28.87
	CM	27.54	27.21	27.49	27.86	28.59	28.61	27.74	27.27	27.19	27.50	28.51	26.83	25.86	26.16
45	IID-HMM	26.30													

Table 1: % Average error rates for various models, each with specific feature dimension, *model type* and $\{l_1,\cdots,l_m\}$.

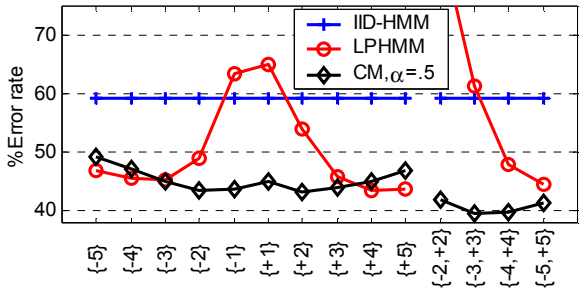

Fig.3: How average error rates varied with the selection of different $\{l_1,\cdots,l_m\}$ for 15-dimension models.

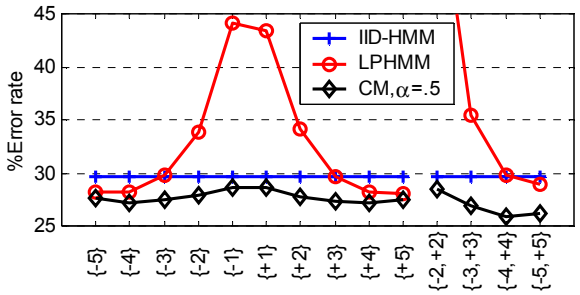

Fig.4: How average error rates varied with the selection of different $\{l_1,\cdots,l_m\}$ for 30-dimension models.

dimension were really worse. Due to the frame overlap in the front-end feature extraction, "pseudo" correlation between frames was introduced and captured, but indiscriminatively, in the above LPHMMs. Moreover, as the inclusion of differentials, the superiority of some LPHMMs with carefully selected $\{l_1,\cdots,l_m\}$ over the traditional HMM diminished, as reported elsewhere in [4]. The dynamics of differentials for different states seemed to be similar to each other and thus less discriminated in LPHMMs, while their statics seemed to be more useful.

The CM was consistently much better than both the baseline and LPHMM in almost all cases, which clearly indicates its superiority over the other two models. The exception that the CMs with $\{-5\}$, $\{-4\}$, $\{+4\}$ and $\{+5\}$ under 15 dimension were slightly worse than the corresponding LPHMM may be attributed to the use of fixed $\alpha = 0.5$. It should be emphasized that the 30-dimension CM with $\{-4,+4\}$ was better than the 45-dimension traditional HMM (25.86% *vs* 26.30%), while both the memory and computation cost were reduced about $1/9 \approx 11\%$.

5.2 Effect of MMI-optimized α

We chose to experiment with the 15-dim CM using $\{-3,+3\}$, which was the best one under 15 dimension in Table 1. Fig. 5 plots the typical changes of the MMI value (6) and the error rate for each iteration, as described in **4.1**. The result clearly demonstrates the effectiveness of the new training procedure. Although the interval bipartition method does not necessarily increase the MMI value for each iteration, it converges quickly, and at the same time the error rate is also reduced to its smallest (38.48%).

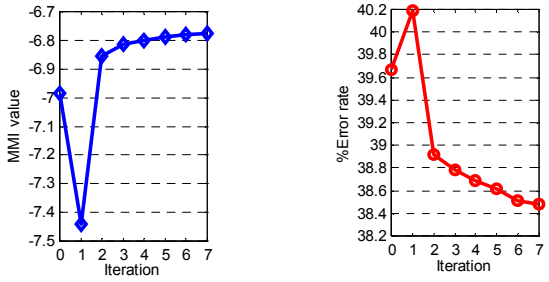

Fig. 5: MMI value and error rate as a function of the iteration index, for the 15-dimension CM with $\{-3,+3\}$.

6. CONCLUSIONS

This paper investigates a recently proposed [7] combined model of statics-dynamics of speech. An optimal combination using MMI is introduced. Experiments on a speaker-independent LVCSR task showed its advantages over both models, with consistent reduction in error rate. Furthermore, the combination weight 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 and J. Li, "Relax frame independence assumption for standard HMMs by state dependent auto-regressive feature models", Proc. ICASSP 2001.

[7] Zhijian Ou and Zuoying Wang, "A new combined model of statics-dynamics of speech", Proc. ICASSP 2002 (to appear).

[8] L.R. Bahl, et al., "Maximum mutual information estimation of hidden Markov model parameters for speech recognition", Proc. ICASSP 1986.