In: **International Conference on Spoken Language Processing (ICSLP)**, Jeju, Korea, 2004,10.

# DISCRIMINATIVE COMBINATION OF MULTIPLE LINEAR PREDICTIONS FOR SPEECH RECOGNITION

*Zhijian Ou, Zuoying Wang*

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

## ABSTRACT

In this paper, new analyses are provided for the problems of applying linear prediction (LP) HMMs in speech recognition. It is shown that, apart from simply aggregating all predictors in one LP, which produces inconsistent results, 'combination' provides another useful way to implement complex dependencies. A method by discriminative combination of multiple LPs (DCoLP) is proposed, with a component-LP selection heuristic. The resulting DCoLP model was tested on a speaker-independent, large vocabulary continuous speech recognition task, and showed improved performance over the standard HMM with comparable computation cost.

## 1. INTRODUCTION

It has long been recognized that the state-conditional independence assumption for the observations in the standard HMM is inaccurate for modeling speech. It neglects the temporal dependencies inherent in the speech signal.

Various methods have been proposed to incorporate temporal dependencies into the standard HMM. One way is to augment original static features with their differentials, and the acoustic model still uses the standard HMM. This method has been shown to be of practical benefit in improving recognition. But it makes a direct violation of the conditional independence assumption. On the other hand, more efforts have been made to study alternative statistical models to the standard HMM. Different degrees of success have been reported. However, the gain from these theoretical models is often limited, in contrast to the effectiveness of the differential feature technique, which is now widely used.

Remarkably, the approach that directly conditions current observation on nearby observations with linear prediction (LP) [1-3] is attractive, since it maintains the efficient Viterbi alignment and decoding. The number of nearby observations used to predict current observation is called the *order* of the LP. Different selections of the (*m*-order) offset-array $L=\{l_1,\cdots,l_m\}$ give different realizations of LPHMMs in practice. Using this notation, the standard HMM could be viewed as a zero-order LPHMM {}.

In previous works [1-3], the selection of the predictor offsets was *arbitrary*. The gain from (arbitrarily) adding predictors (e.g. from {}→{-1}, {-3}→{-3,3}, etc) was small or even often negative. In [3], various 2-order LPHMMs {4,9}, {-4,4} and {3,4} were tested, which were all worse than the 1-order LPHMM {4}, both for the training *and testing* data (suggesting that insufficient training is not the problem). Moreover, in contrast to the standard HMM with differential features, LPHMMs using static features was less effective [1][4]. How to select and use LP dependencies between the observations effectively for speech recognition seems unresolved, which is the main issue addressed in this paper.

First, new analyses of applying LP to exploit temporal dependencies for speech recognition are provided. The analyses are based on a concept of minimum a posterior entropy (MAPE), which in itself is not new and similar to maximum mutual information (MMI) [4], but is more helpful for the analyses here. It is shown that, apart from simply aggregating all predictors in one LP, which produces inconsistent results, 'combination' provides another useful way to *implement* complex dependencies. A method by discriminative combination of multiple LPs (DCoLP) is proposed. The resulting DCoLP model works with log-linear combined multiple component-LPs, which are selected and combined in a principled way guided by MAPE.

After a brief outline of LPHMMs in section 2, we describe in detail the new modeling method in section 3, including the analyses, and comparisons with other works in improving LPHMMs such as discriminative training of LPHMMs [4], other forms of combination [5-7], and Buried Markov Models (BMM) [8]. Experimental results are given in section 4.

## 2. LINEAR PREDICTION HMM

The main characteristic of LPHMMs lies in their definition of the state-observation distributions. For an *m*-order LPHMM with offset-array $L=\{l_1,\cdots,l_m\}$, the probability density function (pdf) for state $s$ and observation $o_t$ is defined conditional on the context as:

$$p(o_t \mid Z_t, q_t = s) = \frac{1}{\sqrt{(2\pi)^D |\Sigma_s|}} \exp\left\{-\frac{1}{2} e_t^T \Sigma_s^{-1} e_t\right\} \quad (1)$$

where $e_t = o_t - \left(\sum_{i=1}^{m} \beta_{s,i} o_{t+l_i} + \mu_s\right)$ is the prediction error, modeled by a single Gaussian distribution with zero mean and full covariance $\Sigma_s$, independent between observations. The LP model predicts current observation $o_t$ using the predictors $Z_t = \{o_{t+l} \mid l \in L\}$. Here $l_i$, $\beta_{s,i} \in R^{D\times D}$ are respectively the offset and the prediction matrix associated with the *i*-th predictor, $\mu_s$ accounts for a constant prediction term.

## 3. DISCRIMINATIVE COMBINATION OF MULTIPLE LPs

An illustrative analysis was given in [9] for the problems of applying LPHMMs in speech recognition. It was shown that a combined model achieved better recognition performance than both the standard HMM and LPHMMs [9][10]. In the following, this idea is extended more rigorously.

### 3.1. Minimum a posteriori entropy (MAPE)

Examination of the effect of adding predictors in terms of some discriminative criterion helps us to gain the insight into the property of LPHMMs in speech recognition.

The a posteriori entropy is defined to measure the remaining uncertainty of the text-message $W$ given the acoustic-observation $O$ with the *estimated* a posteriori distribution $p_\lambda(W|O)$ (parameterized by $\lambda$):

$$H_\lambda(W|O) = E\left[-\log p_\lambda(W|O)\right] \quad (2)$$

The expectation is taken over the *true* distribution $p(W,O)$. In fact, (2) is a conditional entropy-like quantity. Note that

$$H_\lambda(W|O) = E\left[-\log p(W|O)\right] + E\left[\log \frac{p(W|O)}{p_\lambda(W|O)}\right] \quad (3)$$
$$\geq E\left[-\log p(W|O)\right]$$

This says that, the a posteriori entropy with the *estimated* $p_\lambda(W|O)$ is always greater than that with the *true* $p(W|O)$. Only when $p_\lambda(W|O)$ is accurately estimated, i.e. equals the true $p(W|O)$, the a posteriori entropy $H_\lambda(W|O)$ attains its minimum. So $H_\lambda(W|O)$ measures how well the *estimated* $p_\lambda(W|O)$ approximates the *true* $p(W|O)$, and indicates the quality of the implemented plug-in MAP decoder.

In practice, by assuming that the training data, say, $N$ pairs of text-labels and acoustic-observations $(W_n, O_n)$, $n = 1, \cdots, N$, are representative and replacing the expectation by the sample average, $H_\lambda(W|O)$ is given by

$$H_\lambda(W|O) = -\frac{1}{N}\sum_{n=1}^{N} \log p_\lambda(W_n|O_n) \quad (4)$$

Now different from recognizer design based on maximum likelihood (ML), we have a new approach of recognizer design based on MAPE, which aims to find the a posteriori distribution estimator $p_\lambda(W|O)$ directly so as to minimize $H_\lambda(W|O)$. Note that $p_\lambda(W|O)$ could be obtained by

$$p_\lambda(W|O) \stackrel{\Delta}{=} \frac{f_\lambda(O|W)p(W)}{\sum_{W'} f_\lambda(O|W')p(W')} \text{ (still sum to 1)} \quad (5)$$

with acoustic model $f_\lambda(O|W)$ of any forms. The acoustic model $f_\lambda(O|W)$ is of interest only to the extent it is used in (5), and not necessarily to be a distribution. New acoustic models investigated below might not retain the properties of a distribution.

For us, MAPE not only is a parameter estimation criterion (in this sense, similar to MMI), but also more importantly means a new recognizer design approach and is helpful for analyses of different model assumptions.

### 3.2. Analyses in the MAPE framework

Suppose that we have two LPHMMs with predictor offset-array $L_1$ (e.g. {-1}) and $L_2$ (e.g. {-2,-3}) respectively. They exploit LP dependencies between current observation $o_t$ and nearby observations $Z_{t,k} = \{o_{t+l} \,|\, l \in L_k\}$ respectively for $k = 1, 2$. Each LPHMM has its own *estimated* a posteriori distributions $p_{\lambda(L_k)}(W|O)$, parameterized by $\lambda(L_k)$, and the corresponding a posteriori entropy $H_{\lambda(L_k)}(W|O)$.

In previous studies, inclusion of more temporal dependencies is simply by 'aggregation', that is to aggregate all predictors in one LP. In this way, a new LP is built with the offset-array $L_{1\cup 2} = L_1 \cup L_2$ (i.e. {-1,-2,-3} in this example). When the parameters of the new LP, $\lambda(L_{1\cup 2})$ are obtained via ML estimation as in most studies, the a posteriori entropy with the new LP model $H_{\lambda(L_{1\cup 2})}(W|O)$ is not guaranteed to be reduced. Note that

$$\begin{aligned} & H_{\lambda(L_{1\cup 2})}(W|O) \\ &= E\left[-\log p_{\lambda(L_{1\cup 2})}(W|O)\right] \\ &= -E\left[\log p_{\lambda(L_{1\cup 2})}(W,O)\right] + E\left[\log p_{\lambda(L_{1\cup 2})}(O)\right] \\ &= -E\left[\log p_{\lambda(L_{1\cup 2})}(W,O)\right] + E\left[\log \sum_{W'} p_{\lambda(L_{1\cup 2})}(W',O)\right] \end{aligned}$$

ML estimation only guarantees to increase the first term, the log-likelihood of the assumed model. A potential problem is that the implemented dependencies might also increase the likelihood in the context of a different and competing class $W'$. The a posteriori entropy is thus penalized from the second opposite-changing term, which may render the overall change being increased. Hence the recognition performance of the aggregated model $L_{1\cup 2}$ might become worse than before 'aggregation' (i.e. $L_1$, $L_2$). This agrees with the knowledge that a ML-trained model does not necessarily give better discrimination since the model is not the true model.

Here we propose an alternative method to 'aggregation' for adding more predictors. We define a discriminant function by 'combination' of the two LPs as follows ($\gamma_1, \gamma_2 \geq 0$):

$$f\left(o_t \,|\, Z_{t,1\oplus 2}, q_t\right) = p\left(o_t \,|\, Z_{t,1}, q_t\right)^{\gamma_1} p\left(o_t \,|\, Z_{t,2}, q_t\right)^{\gamma_2} \quad (6)$$

where the symbol $\oplus$ denotes the 'combination' operation, different from the conventional 'aggregation' operation $\cup$. The offset-array structure of the new combined model is denoted as $L_{1\oplus 2} = L_1 \oplus L_2$ (i.e. {-1}{-2,-3} in this example). Using (5) and (6), we could obtain the new a posteriori distribution estimator $p_{\lambda(L_{1\oplus 2})}(W|O)$, parameterized by $\lambda(L_{1\oplus 2})$. Notably, it can be shown that the a posteriori entropy $H_{\lambda(L_{1\oplus 2})}(W|O)$ is a *convex* function of the combination weights $\gamma = (\gamma_1, \gamma_2) \geq 0$. Hence there exists one and only one optimum point $\gamma^*$. In this setting, we have

$$H_{\lambda^*(L_{1\oplus 2})}(W|O) \leq \min\left\{H_{\lambda(L_1)}(W|O), H_{\lambda(L_2)}(W|O)\right\} \quad (7)$$

since the right hands are special cases of $H_{\lambda(L_{1\oplus 2})}(W|O)$ with $\gamma = (1,0) / \gamma = (0,1)$. Therefore, the a posteriori entropy with the optimal-weighted combined model is smaller than that with either LP model before 'combination'. The recognition performance of the combined model $L_{1\oplus 2}$ is most probable to be improved.

### 3.3. Formulation

Consider an acoustic model where the discriminant function

concerning the observations $O = o_1 \cdots o_T$ given the state sequence $Q = q_1 \cdots q_T$ is defined as:

$$f(O|Q) = \prod_{t=1}^{T} f(o_t | Z_t, q_t) \quad (8)$$

where $f(o_t | Z_t, q_t)$ is the state-observation discriminant function for the state $q_t$ and observation $o_t$, depending on $Z_t$; $Z_t \subset \{\cdots o_{t-1} o_{t+1} \cdots\}$ is a subset of $o_t$'s surrounding context to be used as predictors. Introducing discriminative combination of multiple LPs (DCoLP), we define

$$f(o_t | Z_t; q_t) = \prod_{k=1}^{K} p(o_t | Z_{t,k}; q_t)^{\gamma_{s,k}} \quad (9)$$

where $p(o_t | Z_{t,k}; q_t)$ is the conditional pdf computed by the *k*-th component-LP with offset-array $L_k$, which captures the LP dependency between $o_t$ and the predictors $Z_{t,k} = \{o_{t+l} | l \in L_k\}$. $\gamma_{s,k} \geq 0$ are the state-specific weights, or as $\gamma_k$ when tied globally across all states ($1 \leq k \leq K$). Denote the combination weights by $\gamma$ as a whole.

The important property is that, the a posteriori entropy $H_{\lambda(L_{1\oplus\cdots\oplus K})}(W|O)$ is a *convex* function of the combination weights $\gamma$. Thus, similar to the arguments in section 3.2, the a posteriori entropy with the optimal-weighted DCoLP model will be smaller than that with each component-LP model:

$$H_{\lambda^*(L_{1\oplus\cdots\oplus K})}(W|O) \leq \min_{1\leq k\leq K} H_{\lambda(L_k)}(W|O) \quad (10)$$

There are two parts of parameters in a DCoLP model. 1) The combination weights are trained under MAPE criterion. Gradient descent method with line search using interval bi-partition performs efficiently to reach the global minimum for this convex optimization. 2) The LP parameters are currently subject to ML training. Once statistics are gathered, model parameters of the $K$ component-LPs are re-estimated separately. The overall alternate training procedure is analogous to [10].

The resulting DCoLP model is characterized and will be denoted by its particular offset-array structure, $L_{1\oplus\cdots\oplus K} = L_1 \oplus \cdots \oplus L_K$. Through reasonable 'aggregation' and 'combination', the resulting DCoLP model will try to produce the best (measured by the MAPE criterion) dependency structure from the predictor pool, providing the best discrimination. Intuitively, when 'combination' is more beneficial than 'aggregation' for increasing the discrimination, one should use 'combination'. Otherwise, it is reasonable to use 'aggregation'. It is clear that we need some structure learning method, rather than by crude search or arbitrary choose. This is achieved by *a component-LP selection heuristic.*

The MAPE-trained combination weights reflect the discrimination ability of the corresponding component-LPs. They could be used to guide the selection of which LPs to use. We could start from an initial DCoLP model, which includes the predictors of interest simply as 1-order LPs (e.g. {-2}{2}{-4}{4}{-6}{6}). Then step-by-step, we could cancel small-weighted LPs and aggregate predictors when beneficial. Finally, a compact and discriminative offset-array structure will be found. In section 4, this method is experimentally studied and works well.

### 3.4. Discussion

In view of the shortcoming of ML estimation, it has been tried to train the LP parameters under some discriminative criterion [4]. If the parameters $\lambda(L_{1\cup 2})$ are obtained by MAPE training, initialized from $\lambda(L_1)$ or $\lambda(L_2)$, theoretically $H_{\lambda(L_{1\cup 2})}(W|O)$ will become smaller than $H_{\lambda(L_1)}(W|O)$ or $H_{\lambda(L_2)}(W|O)$. However, for discriminative (MMI or MAPE) training, the objective function is not a convex function of the LP parameters. There does not exist any optimization method guaranteed with fast and reliable convergence. Therefore in practice the improvement is usually limited by the hard and complex optimization procedure. In [4], when applying MMI training, LPHMM {-7,-5,-3} did slightly better than {-5}, and was still worse than the standard HMM with differential features. In contrast, the proposed DCoLP approach is simpler, and as we'll see in the experiments, is more effective.

Different from the log-linear combination (LLC) used in (9), there exist other forms of combination of multiple conditional pdf's from LPs, including linear opinion pool (LiOP) [5-6] and logarithmic opinion pool (LgOP) [7]. However, when considering both-sided dependencies [6], the two LiOP-derived one-sided pdf's were still combined using LLC; and the offset-array, originally to be optimized, was fixed for practical reasons. In [7], LLC was also introduced to improve the discrimination. Furthermore remarkably, MAPE training of the combination weights in LLC is a kind of *convex* optimization, while this property is not observed in any other forms of combination (LiOP, LgOP).

Finally, DCoLP essentially says that, apart from simple 'aggregation' of all parent variables in one LP, 'combination' provides another useful way to *implement* complex dependencies. In this sense, it is orthogonal to the BMM approach [8], which is mainly to find *which dependencies* are to be included (helpful for discrimination), using a heuristic, pairwise selection algorithm. Given the selected dependency structure, there might have various ways to implement the dependencies between the variables, including DCoLP, etc. Furthermore, the DCoLP approach itself may provide a way to find the discriminative dependency structure, with the component-LP selection heuristic describe above.

## 4. EXPERIMENTAL RESULTS

Experiments were carried on a speaker-independent Chinese LVCSR task using the male speech database for "China National 863 Assessment". Utterances from 76 speakers were used for training and those from the other 7 speakers for testing (with about 600 sentences for each speaker), of all the models described below. The system used 100 consonant units each with 2 states, 164 vowel units each with 4 states, plus one single-state silence model. Here we focus on the *acoustic part* of the recognition system [10], and report the first-candidate syllable error rate (SER).

The speech was parameterized into 14 MFCCs along with the normalized log-energy, and their first and second order differentials. Results using full-predictors were reported here. The corresponding diagonal-predictor versions were also trained and used as seeds.

Table 1-4 shows our study to build a high-performance DCoLP model using only the 15-dimention static features. The

Table 1. %SER for optimal-global-weighted DCoLP models, using only the 15-dim static features in the process of the component-LP selection. The superscripts are the resulting optimized weights, displayed as $\bar{\gamma}_k = \gamma_k / \sum \gamma_{k'}$.

| | | |
|---|---|---|
| $\{-2\}^{0.32}\{2\}^{0.25}\{-4\}^{0.06}\{4\}^{0.10}\{-6\}^{0.13}\{6\}^{0.14}$ | 27.83% | cancel {-4} |
| $\{-2\}^{0.35}\{2\}^{0.26}\{4\}^{0.09}\{-6\}^{0.17}\{6\}^{0.13}$ | 27.86% | aggregation {-6,6} |
| $\{-2\}^{0.32}\{2\}^{0.21}\{4\}^{0.00}\{-6,6\}^{0.47}$ | 26.65% | cancel {4} |
| $\{-2\}^{0.30}\{2\}^{0.23}\{-6,6\}^{0.47}$ | 26.61% | |

Table 2. %SER for related LPHMMs

| | |
|---|---|
| {-2,-6,6} | 35.64% |
| {-6,-4,-2} | 37.82% |
| {-2,2,-6,6} | 62.39% |

component-LP selection heuristic was taken. The initial DCoLP model was built as {-2}{2}{-4}{4}{-6}{6}. Dependencies were allowed to span a maximum of 60ms (6 frames) on either side of *t*. In view of the frame overlap in the front-end feature extraction, the predictors of interest were selected with one frame apart to avoid redundant modeling.

Table 1 shows how the dependency structure was discriminatively optimized step by step. By canceling small-weighted {-4}, beneficial aggregation of {-6,6} and canceling {4}, the resulting DCoLP model {-2}{2}{-6,6} was more compact and discriminative, with the error rate of 26.61%. Simply aggregation of the four predictors to use {-2,2,-6,6} gave the error rate of 62.39%, which was much larger. The error rates for related LPHMMs are given in Table 2 for comparison. Through learning reasonable 'aggregation' and 'combination', the resulting DCoLP model was the best among those models to discriminatively exploit the temporal dependencies for speech recognition.

We experimented with different models using two frames chosen at offsets -2, -4, -6, 2, 4, 6. Various 2-order LPs and their counterpart DCoLP models (combination of two 1-order LPs) were compared in Table 3. In all test cases, the DCoLP model outperformed the corresponding 2-order LP, except that {-6,6} was better than {-6}{6}. Thus {-6,6} replaced {-6}{6} as a component-LP. Note that further adding predictors to {-6,6} by aggregation failed, as can be seen from Table 2. But by combination, we could still utilize more predictors.

Starting from the optimal-global-weighted model {-2}{2}{-6,6}, the corresponding state-specific-weighted model was built, which is shown in Table 4 to further reduce the error rate. The optimal-state-specific-weighted DCoLP model {-2}{2}{-6,6} using only the 15-dimention static features outperformed the standard HMM using 45-dimention features (+Δ+ΔΔ), with comparable computation cost. The error rate reduction was 6% (from 26.30% to 24.74%).

Here for comparison with the standard HMM, the sum of the state-specific weights, $\sum_k \gamma_{s,k}$ was constrained to the same for all states, being equal to $\sum_k \gamma_k$, the sum of the previously trained optimal global-weights.

## 5. CONCLUSIONS

In this paper, a new method by discriminative combination of multiple LPs (DCoLP) is introduced, which provides another useful way to implement complex LP dependencies, in view of the inconsistent results of 'aggregation'. The convexity property herein ensures that the combination weights are efficiently optimized, and the a posteriori entropy is effectively reduced compared with each component-LP model. Using a component-LP selection heuristic, the resulting DCoLP model was tested on a speaker-independent LVCSR task and showed improved performance over the standard HMM with comparable computation cost.

Table 3. Comparison (%SER) of various 2-order LPHMMs and their counterpart DCoLP models (combination of two 1-order LPs), using two frames chosen at offsets -2, -4, -6, 2, 4, 6.

| | | | |
|---|---|---|---|
| {-2,2} | {-2}{2} | {-4,4} | {-4}{4} |
| 62.83% | 31.58% | 35.76% | 30.35% |
| {-2,4} | {-2}{4} | {-4,6} | {-4}{6} |
| 43.54% | 29.19% | 32.34% | 31.68% |
| {-2,6} | {-2}{6} | {-4,-6} | {-4}{-6} |
| 37.50% | 29.72% | 35.41% | 34.27% |
| {-2,-4} | {-2}{-4} | {-6,6} | {-6}{6} |
| 39.14% | 34.08% | 32.90% | 33.67% |
| {-2,-6} | {-2}{-6} | | |
| 38.11% | 33.32% | | |

Table 4. Comparison of the DCoLP model {-2}{2}{-6,6} using only the 15-dim static features and the standard HMM using 45-dim features (+Δ+ΔΔ) in terms of %SER and Number of multiplicaiton(*)/addition(+) per state-observation model computation.

| | | %SER | Num of */+ |
|---|---|---|---|
| 15-dim {-2}{2}{-6,6} | global-weighted | 26.61 | 1347/1353 |
| | state-specific-weighted | 24.74 | |
| HMM, 45-dim (+Δ+ΔΔ) | | 26.30 | 1124/1126 |

## 6. REFERENCES


[1] 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.

[2] P.C. Woodland, "Hidden Markov models using vector linear prediction and discriminative distributions", in *ICASSP* 92.

[3] B.A. Maxwell and P.C. Woodland, "Hidden Markov models using shared and global vector linear predictiors", in *Eurospeech* 1993.

[4] K.K. Chin and P.C. Woodland, "Maximum mutual information training of hidden Markov models with vector linear predictors", in *ICSLP* 2002.

[5] J. Ming and F.J. Smith, "Modeling of the inter-frame dependence in an HMM using conditional gaussian mixtures", *Computer Speech and Language*, Vol.10, 1996.

[6] P. Hanna, et al., "Modeling inter-frame dependence with proceeding and succeeding frames", in *Eurospeech* 1997.

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

[8] J.A. Bilmes, "Buried Markov models for speech recognition", in *ICASSP* 99.

[9] Zhijian Ou and Zuoying Wang, "A new combined model of statics-dynamics of speech", in *ICASSP* 2002.

[10] Zhijian Ou and Zuoying Wang, "A combined model of statics-dynamics of speech optimized using maximum mutual information", in *ICSLP* 2002.