Discriminative Combination of Multiple Linear Predictions for Speech Recognition

Zhijian Ou

Department of Electronic Engineering
Tsinghua University
Beijing, China

Email: ozj@tsinghua.edu.cn

Abstract

- New analyses are provided for the problems of applying LPHMMs in speech recognition.
  - 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.
  - Works with log-linear combined multiple component-LPs.
- Showed improved performance over the standard HMM with comparable computation cost.

Discriminative Combination of Multiple Linear Predictions for Speech Recognition

1. Introduction
2. Analyses
3. Model formulation
4. Experimental results

Introduction

- The state-conditional independence assumption for the observations in the standard HMM is inaccurate for modeling speech.

Introduction

- One way is to augment original static features with their differentials, and the acoustic model still uses the standard HMM.
  - Makes a direct violation of the conditional independence assumption.
- More efforts have been made to study alternative statistical models to the standard HMM.
  - The gain from these theoretical models is often limited, in contrast to the effectiveness of the differential feature technique, which is now widely used.

Linear Prediciton HMM

- Directly condition current observation on nearby observations with linear prediction (LP)
  - Attractive
    - Maintain the efficient Viterbi alignment and decoding.
  - For an m-order LPHMM with offset-array $\{l_1, \ldots, l_m\}$, the pdf for state s and observation o_t is defined conditional on Z_t as:

$$p(o_t \mid Z_t, q_t = s) = \frac{1}{\sqrt{(2\pi)^D |\Sigma_s|}} \exp\left\{-\frac{1}{2} e_t^T \Sigma_s^{-1} e_t\right\} \qquad Z_t = \{o_{t+l} \mid l \in L\}$$

    - $e_t = o_t - \left(\sum_{i=1}^{m} \beta_{s,i} o_{t+l_i} + \mu_s\right)$ is the prediciton error $\sim N(0, \Sigma_s)$
  - Different selections of the (m-order) offset-array $\{l_1, \ldots, l_m\}$ give different realizations of LPHMMs in practice.

Problem of applying LP

- The selection of the predictor offsets was *arbitrary*.
  - The gain from (arbitrarily) adding predictors (e.g. from $\{\} \rightarrow \{-1\}$, $\{-3\} \rightarrow \{-3,3\}$, etc) was small or even often negative.
  - In contrast to the standard HMM with differential features, LPHMMs using static features was less effective.
- 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.

Discriminative Combination of Multiple Linear Predictions for Speech Recognition

1. Introduction
2. Analyses
3. Model formulation
4. Experimental results

Minimum a posteriori entropy (MAPE)

- MAPE in itself is not new and similar to cross-entropy, but is more helpful for the analyses here.

Minimum a posteriori entropy (MAPE)

- The a posteriori entropy with the *estimated* a posteriori distribution $p_\lambda(W \mid O)$ (parameterized by λ):

$$
\begin{aligned}
H_\lambda(W \mid O) &= E\left[-\log p_\lambda(W \mid O)\right] \\
&= E\left[-\log p(W \mid O)\right] + E\left[\log \frac{p(W \mid O)}{p_\lambda(W \mid O)}\right] \\
&\geq E\left[-\log p(W \mid O)\right]
\end{aligned}
$$

  - $H_\lambda(W \mid O)$ measures how well the *estimated* $p_\lambda(W \mid O)$ approximates the *true* $p(W \mid O)$, and indicates the quality of the implemented plug-in MAP decoder.

Recognizer design based on MAPE

- A new approach of recognizer design based on MAPE
  - Aims to find the a posteriori distribution estimator $p_\lambda(W \mid O)$ directly so as to minimize $H_\lambda(W \mid O)$.
  - Note that $p_\lambda(W \mid O)$ could be obtained by

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

with acoustic model $f_\lambda(W \mid O)$ of any forms.

- New acoustic models investigated below might not retain the properties of a distribution.

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.

Analyses in the MAPE framework

- In previous studies, inclusion of more temporal dependencies is simply by 'aggregation'.

  - 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 $H_{\lambda(L_{1\cup 2})}(W \mid O)$ is not guaranteed to be reduced.

  - Hence the recognition performance of the aggregated model $L_{1\cup 2}$ might become worse than before 'aggregation' (i.e. L_1, L_2)

An alternative method

- Propose an alternative method to 'aggregation' for adding more predictors
  - Define a discriminant function by 'combination' of the two LPs:

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

  - 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)

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

  - We could obtain the new a posteriori distribution estimator $p_{\lambda(L_{1\oplus 2})}(W \mid O)$, parameterized by $\lambda(L_{1\oplus 2})$.

An alternative method

- It can be shown that the a posteriori entropy $H_{\lambda(L_{1\oplus 2})}(W \mid O)$ is a *convex* function of the combination weights $\gamma = (\gamma_1, \gamma_2) \geq 0$.

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

- The a posteriori entropy with the optimal-weighted combined model is smaller than that with either LP model before 'combination'.

Discriminative Combination of Multiple Linear Predictions for Speech Recognition

1. Introduction
2. Analyses
3. Model formulation
4. Experimental results

Model Formulation

- Introducing discriminative combination of multiple LPs (DCoLP), we define

$$f\left(o_t \mid Z_t; q_t\right) = \prod_{k=1}^{K} p\left(o_t \mid Z_{t,k}; q_t\right)^{\gamma_{s,k}}$$

  - $p(o_t \mid Z_{t,k}, q_t)$ 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 $Z_{t,k}=\{o_{t+l} \mid l \in L_k\}$.
  - $\gamma_{s,k}$ are the state-specific weights, or as γ_s when tied globally across all states.

Property

- The key property is that, the a posteriori entropy $H_{\lambda(L_{1\oplus\cdots\oplus K})}(W \mid O)$ is a *convex* function of the combination weights .
- The a posteriori entropy with the optimal-weighted DCoLP model will be smaller than that with each component-LP model:

$$H_{\lambda^*(L_{1\oplus\cdots\oplus K})}\left(W \mid O\right) \le \min_{1 \le k \le K} H_{\lambda(L_k)}\left(W \mid O\right)$$

A component-LP selection heuristic

- The MAPE-trained combination weights reflect the discrimination ability of the corresponding component-LPs.
  - 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.
  - This method is experimentally studied and works well.

Discussion - 1

- Discriminative training of LPHMMs
  - The objective function is not a convex function of the LP parameters.
  - In practice the improvement is usually limited by the hard and complex optimization procedure.
  - In contrast, the proposed DCoLP approach is simpler, and as we'll see in the experiments, is more effective.

Discussion - 2

- Other forms of combination of multiple conditional pdf's from LPs.
  - DCoLP: Log-linear combination (LLC)
  - For models using linear opinion pool (LiOP) or logarithmic opinion pool (LgOP), LLC was practically introduced to improve the discrimination.
  - 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).

Discussion - 3

- 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, which is mainly to find which dependencies are to be included using a heuristic, pairwise selection algorithm.
  - The DCoLP approach itself may provide a way to find the discriminative dependency structure, with the component-LP selection heuristic describe above.

Discriminative Combination of Multiple Linear Predictions for Speech Recognition

1. Introduction
2. Analyses
3. Model formulation
4. Experimental results

Configuation

- A speaker-independent Chinese LVCSR task
  - Utterances from 76 speakers for training
  - Utterances from the other 7 speakers for testing
  - About 600 sentences for each speaker
- Used 100 consonant units each with 2 states, 164 vowel units each with 4 states, plus one single-state silence model.
- Report the syllable error rate (SER).
- 15-dim static feature: 14 MFCCs + E

Results

%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

$\{-2\}^{0.32}\{2\}^{0.25}\{-4\}^{0.06}\{4\}^{0.10}\{-6\}^{0.13}\{6\}^{0.14}$	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%	

%SER for related LPHMMs

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

Results

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%		

Results

%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

$\{-2\}^{0.32}\{2\}^{0.25}\{-4\}^{0.06}\{4\}^{0.10}\{-6\}^{0.13}\{6\}^{0.14}$	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%	

%SER for related LPHMMs

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

Results

Comparison of the DCoLP model {-2}{2}{-6,6} using only the 15-dim static features and the standard HMM using 45-dim features ($+\Delta+\Delta\Delta$) 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 ($+\Delta+\Delta\Delta$)		26.30	1124/1126

Conclusion

- A new method by discriminative combination of multiple LPs (DCoLP) is introduced.
  - 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.
  - Showed improved performance over the standard HMM with comparable computation cost.

Discriminative Combination of Multiple Linear Predictions for Speech Recognition

Thank you!