Nonreciprocal Data Sharing in Estimating HMM Parameters

Xiaoqiang Luo and *Frederick Jelinek*
Center for Language and Speech Processing (CLSP)
Department of Electrical & Computer Engineering
The Johns Hopkins University
Baltimore, MD21218, USA

ABSTRACT

Parameter tying is often used in large vocabulary continuous speech recognition (LVCSR) systems to balance the model resolution and generalizability. However, one consequence of tying is that the differences among tied constructs are ignored. Parameter tying can be alternatively viewed as reciprocal data sharing in that a tied construct uses data associated with all others in its tied-class. To capture the fine difference among tied HMM constructs, we propose to use nonreciprocal data sharing (NRDS) when estimating HMM parameters. In particular, when estimating Gaussian parameters for a HMM state, contributions from other acoustically similar HMM states will be weighted, thus allowing different statistics to govern different states. Data sharing weights are optimized using cross-validation. It can be shown that the objective function for cross-validation is a sum of rational functions and can be efficiently optimized by the growth-transform [5, 7]. Our results on Switchboard [4] show that NRDS reduces the word error rate (WER) significantly compared with a state-of-art baseline system using HMM state-tying.

1 Basic Idea

One way to think of parameter-tying is that data associated with tied constructs is shared reciprocally. For instance, after two triphone states "A" and "B" are tied, all data of "A" contributes to model "B", and vice versa, thus making the two states effectively identical. Tying brings about "reciprocal" data sharing in the sense that at each Baum-Welch iteration, "A" uses the totality of data contributed by "B", and so does "B" use all the contribution from "A". While this helps to improve model robustness, it at the same time ignores the distinction between "A" and "B". So "non-reciprocal" sharing might be a better choice. In particular, a data sharing factor w(B,A) can be introduced to weight the contribution from state "B" when estimating the Gaussian parameters of state "A", and an independent factor w(A,B) to weight the contribution from state "A" to state "B". In general, final models for "A" and "B" will then be different. Similar approach has been used in analyzing the nonhomogeneity of images [9].

2 Non-Reciprocal Data Sharing

To elaborate the above idea, let's start from the Baum-Welch [1] reestimation of HMM parameters. Without loss of generality, assume that HMM states $\mathcal{S}$ are clustered into disjoint classes $\mathcal{K} = \{\mathcal{C}_1, \mathcal{C}_2, \cdots, \mathcal{C}_K\}$. HMM states belonging to a class share a single set of parameters. The EM [2] auxiliary function for a state-tied system, after ignoring contributions of HMM transition probabilities, is

$$Q_{T_1}(\theta|\theta') \approx \sum_{i=1}^{K} \sum_{s \in \mathcal{C}_i} \sum_{t} \gamma_t(s) \log P(o_t|s; \theta_{C_i}). \quad (1)$$

where $O = \{o_t\}_{t=1}^{T}$ is the training speech, and $\gamma_t(u) = P(S_t = u|O)$ is the occupancy count for state u at time t. θ' and θ are the totality of model parameters before and after an iteration respectively. When necessary, we will use θ_s or $\theta_{\mathcal{C}_i}$ to denote the parameters specific to state s or class $\mathcal{C}_i$. So $\mu_s = \mu_{\mathcal{C}_i}$ and $\Sigma_s = \Sigma_{\mathcal{C}_i}$ for all $s \in \mathcal{C}_i$ in the paradigm of state-tying. If a single Gaussian is assumed as the output distribution, Baum-Welch reestimation for class means and variances are

$$\mu_{\mathcal{C}_i} = \frac{\sum_{u \in \mathcal{C}_i} \sum_{t=1}^{T} \gamma_t(u) o_t}{\sum_{u \in \mathcal{C}_i} \sum_{t=1}^{T} \gamma_t(u)} \quad (2)$$

$$\Sigma_{\mathcal{C}_i} = \frac{\sum_{u \in \mathcal{C}_i} \sum_{t=1}^{T} \gamma_t(u)(o_t - \mu_{\mathcal{C}_i})(o_t - \mu_{\mathcal{C}_i})'}{\sum_{u \in \mathcal{C}_i} \sum_{t=1}^{T} \gamma_t(u)} \quad (3)$$

Notice that the data sharing is indeed reciprocal in the above formulas: for any states u and v in $\mathcal{C}_i$, u uses the totality of data contributed by v, and vice versa. Therefore, the distinction between individual states vanishes. However, this can be improved without sacrificing robustness.

Let $C : \mathcal{S} \to \mathcal{K}$ be the map from HMM states to classes, and denote the HMM class for state s by $C(s)$. Observe

that (1) is equivalent to

$$Q_{T_2}(\theta|\theta') = \sum_s \sum_{u\in C(s)} \sum_t \gamma_t(u) \log P(o_t|s;\theta_s) \quad (4)$$

$$\text{and } \theta'_s = \theta'_{s'},\ if\ C(s) = C(s') \quad (5)$$

in that maximizing (4) under the condition (5) yields the same result as maximizing (1). However, the change of expression provides us with an alternative view of parameter tying: tying a set of parameters is equivalent to first relaxing the constraints on parameters, and then estimating HMM parameters for an *individual* HMM state s by using contributions from all HMM states in $C(s)$. Notice that if $\theta'_s = \theta'_{s'}$ for any s, s' such that $C(s) = C(s')$, then it remains true that $\theta_s = \theta_{s'}$ after an iteration. So parameter tying is equivalent to *data sharing* under the condition (5).

Now we can define a different objective function by weighting contributions of other states:

$$Q_N(\theta|\theta') = \sum_s \sum_{u\in B(s)} \sum_t w(u,s)\gamma_t(u) \log P(o_t|s;\theta_s) \quad (6)$$

where $B(s)$ is a set of HMM states "similar" to s that make contributions to estimating θ_s. Maximizing (6) results in the NRDS update formulae (7) and (8):

$$\bar{\mu}_s = \frac{\sum_u \sum_{t=1}^T w(u,s)\gamma_t(u)o_t}{\sum_u \sum_{t=1}^T w(u,s)\gamma_t(u)} \quad (7)$$

$$\bar{\Sigma}_s = \frac{\sum_u \sum_{t=1}^T w(u,s)\gamma_t(u)(o_t-\bar{\mu}_s)(o_t-\bar{\mu}_s)'}{\sum_u \sum_{t=1}^T w(u,s)\gamma_t(u)} \quad (8)$$

NRDS estimates $\{\bar{\mu}_s, \bar{\Sigma}_s\}$ can be regarded as "smoothing" MLE of an unconstrained (i.e, untied) HMM system. Let $\hat{\mu}_s$ and $\hat{\Sigma}_s$ be the maximum likelihood estimate (MLE) of Gaussian mean and covariance for state s, where

$$\hat{\mu}_s = \frac{\sum_t \gamma_t(s)o_t}{\sum_t \gamma_t(s)} \quad (9)$$

$$\hat{\Sigma}_s = \frac{\sum_t \gamma_t(s)(o_t-\hat{\mu}_s)(o_t-\hat{\mu}_s)'}{\sum_t \gamma_t(s)}. \quad (10)$$

Then NRDS estimates (7) and (8) can be expressed as a function of $\hat{\mu}_s$ and $\hat{\Sigma}_s$.

$$\bar{\mu}_s = \frac{\sum_u w(u,s)\gamma_u\hat{\mu}_u}{\sum_u w(u,s)\gamma_u} \quad (11)$$

$$\bar{\Sigma}_s = \frac{\sum_u w(u,s)\gamma_u(\hat{\Sigma}_u+\hat{\mu}_u\hat{\mu}'_u)}{\sum_u w(u,s)\gamma_u} - \bar{\mu}_s\bar{\mu}'_s \quad (12)$$

where $\gamma_u = \sum_{t=1}^T \gamma_t(u)$.

The interpretation of (11) and (12) is that $\hat{\mu}_s$ and $\hat{\Sigma}_s$ are reestimated for states with positive occupancy counts. Then for states with insufficient data, (11) and (12) are carried out to get reliable estimates by smoothing MLE $\{\hat{\mu}_s, \hat{\Sigma}_s\}$. Since NRDS is not MLE, the likelihood of training data does not necessarily increase after an iteration. However, under some conditions, upper and lower bounds can be established for NRDS estimates. For details, readers are referred to [8].

In the above discussions, a single Gaussian is assumed to be the state output distribution. Extension to HMMs with a mixture of Gaussians as the state output distribution is straightforward if data-sharing is to be carried out at the mixture component level. However, doing so will result in a large number of sharing weights. As will be shown shortly, data sharing weights will be estimated from training data as well. Too many weights will make it difficult to get reliable weight estimates. Therefore, we insist that data sharing be carried out at the HMM state level. We first weigh occupancy count of state u at time t by $w(u,s)$, and then distribute the weighted counts to the state s based on how likely a mixture component generates a speech frame. That is,

$$\eta_t(u,s.m) = \gamma_t(u)\frac{c_{s.m}N(o_t|s.m)}{\sum_{i=1}^{M_s} c_{s.i}N(o_t|s.i)} \quad (13)$$

$$\bar{\gamma}_t(s.m) = \sum_{u\in B(s)} w(u,s)\eta_t(u,s.m) \quad (14)$$

where $s.m$ stands for the m^{th} mixture component of state s, M_s is the number of mixture components of state s, and $\eta_t(u,s.m)$ is the occupancy count that state u contributes to the m^{th} component of state s at time t. $c_{s.m}$ is the mixture weight of $s.m$, and $N(\cdot|s.m)$ is the m^{th} Gaussian distribution of state s. With this notation, the NRDS mean for the m^{th} component of state s can be written as

$$\bar{\mu}_{s.m} = \frac{\sum_{u\in B(s)} w(u,s)\sum_t \eta_t(u,s.m)o_t}{\sum_{u\in B(s)} w(u,s)\sum_t \eta_t(u,s.m)}. \quad (15)$$

Reestimation formulas for Gaussian covariances can be established similarly [8].

3 Optimizing Data Sharing Weights

Since sharing weights can be regarded as smoothing the MLE estimate, directly optimizing the likelihood of training data will yield trivial W, that is, zero weights for cross states. Therefore, it is necessary to optimize W over an independent set of data. To this end, cross-validation or deleted-interpolation [6] will be adopted to find the optimal W. The procedure can be outlined as follows.

Let the training data O be partitioned disjointly into $O^{(1)}, O^{(2)}, \cdots, O^{(P)}$, and let $O^{(\bar{p})} = O - O^{(p)}$ for $p = 1, 2, \cdots, P$. We will start with an initial model $\mathcal{M}_0$ and get an estimate of HMM parameters $\mathcal{M}^{(\bar{p})} = \{\bar{\mu}_s^{(\bar{p})}, \bar{\Sigma}_s^{(\bar{p})}\}$ out of $O^{(\bar{p})}$ using non-reciprocal data sharing. Hence

$\mathcal{M}^{(\bar{p})}$ is a function of data sharing weights W. Then we will evaluate $Q(O^{(p)}; \mathcal{M}^{(\bar{p})})$, the EM auxiliary function[1] on data $O^{(p)}$ using the model $\mathcal{M}^{(\bar{p})}$. The "optimal" W^* will be the one that maximizes $\sum_{p=1}^{P} Q(O^{(p)}; \mathcal{M}^{(\bar{p})})$, or

$$W^* = \arg\max_W Q(W) = \arg\max_W \sum_{p=1}^{P} Q(O^{(p)}; \mathcal{M}^{(\bar{p})}).$$

To facilitate the derivation of the above objective function, the following notation or conventions will be adopted. Let $N_s = |B(s)|$ be the cardinality of the candidate set $B(s)$ for state s, and when it is necessary to enumerate HMM states in $B(s)$, we will write $B(s)$ as

$$B(s) = \{s_1, s_2, \cdots, s_{N_s}\}. \quad (16)$$

Weights $\{w(s_i, s)\}$ will be written as a vector w_s when necessary. In addition, a $\hat{}$ and a $\bar{}$ on top of a symbol denote MLE and NRDS estimates, respectively. A superscript p indicates that a quantity is obtained from or depends on data partition $O^{(p)}$ while a superscript $\bar{p}$ means a quantity is associated with $O^{(\bar{p})} = O - O^{(p)}$, all the data except the segment $O^{(p)}$. If the time subscript t is dropped in occupancy counts, it means it has been summed over the time index, i.e, $\gamma^p(s.m) = \sum_t \gamma_t^p(s.m)$.

$\sum_p Q(O^{(p)}; \mathcal{M}^{(\bar{p})})$ can be written as a sum of functions each of which depends on weights related to only one state:

$$\sum_p Q(O^{(p)}; \mathcal{M}^{(\bar{p})}) = \sum_s Q_s(w_s) \quad (17)$$

where $Q_s(w_s)$ is

$$Q_s(w_s) = \sum_{p=1} \sum_t \gamma_t^{(p)}(s.m) \log N(o_t^{(p)}|s.m). \quad (18)$$

The Gaussian parameters in $N(o_t^{(p)}|s.m)$ are obtained using NRDS on data $O^{(\bar{p})}$, and are therefore functions of data sharing weights w_s. When optimizing w_s, we will assume that Gaussian covariances are known. But they will be updated once the optimal weights w_s are available.

Define

$$\nu_{u,s.m}^{(\bar{p})} = \frac{\sum_t \eta_t^{(\bar{p})}(u, s.m) o_t^{(\bar{p})}}{\sum_t \eta_t^{(\bar{p})}(u, s.m)} \quad (19)$$

$$G_{s.m}^{(\bar{p})} = [\eta^{(\bar{p})}(s_1, s.m)\nu_{s_1,s.m}^{(\bar{p})}, \cdots, \eta^{(\bar{p})}(s_{N_s}, s.m)\nu_{s_{N_s},s.m}^{(\bar{p})}] \quad (20)$$

$$g_{s.m}^{(\bar{p})} = [\eta^{(\bar{p})}(s_1, s.m), \cdots, \eta^{(\bar{p})}(s_{N_s}, s.m)]', \quad (21)$$

Note that $G_{s.m}^{(\bar{p})}$ is an $n \times N_s$ matrix while $g_{s.m}^{(\bar{p})}$ is an N_s-dimensional vector. Therefore, NRDS mean obtained from data $O^{(\bar{p})}$ is

$$\bar{\mu}_{s.m}^{(\bar{p})} = \frac{G_{s.m}^{(\bar{p})} w_s}{g_{s.m}^{(\bar{p})'} w_s}. \quad (22)$$

[1] EM counts are obatined from the initial model $\mathcal{M}_0$

Let

$$\hat{\mu}_{s.m}^{(p)} = \frac{\sum_t \gamma_t^{(p)}(s.m) o_t^{(p)}}{\sum_t \gamma_t^{(p)}(s.m)} \quad (23)$$

be the MLE mean of the m^{th} component of state s. Plug (22) into (18), and discard terms independent of w_s, we obtain

$$Q_s(w_s) = \frac{1}{2} \sum_{p=1}^{P} \sum_{m=1}^{M_s} \frac{w_s' A_{p,m}(s) w_s}{w_s' D_{p,m}(s) w_s}, \quad (24)$$

where

$$A_{p,m}(s) = \gamma^{(p)}(s.m)\big(2g_{s.m}^{(\bar{p})}\hat{\mu}_{s.m}^{(p)'}\bar{\Sigma}_{s.m}^{-1}G_{s.m}^{(\bar{p})} - G_{s.m}^{(\bar{p})'}\bar{\Sigma}_{s.m}^{-1}G_{s.m}^{(\bar{p})}\big) \quad (25)$$

$$D_{p,m}(s) = g_{s.m}^{(\bar{p})}g_{s.m}^{(\bar{p})'}. \quad (26)$$

The objective function (24) is a sum of rational functions, which can be maximized efficiently by the growth transform [5, 7]. Details of the optimization algorithm can be found in a companion paper in SST-98 [7].

4 Experimental Results

We implemented the proposed non-reciprocal data sharing and tested it on the Switchboard (SWBD) [4] task. Since NRDS is developed by generalizing the idea of tying HMM states, NRDS results are compared with those achieved by a state-tied baseline system. The baseline system was built in the LVCSR workshop WS97 [3].

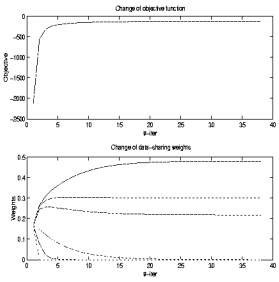

Figure 1: Top: objective function vs. iterations; Bottom: data-sharing weights vs. iterations

The WS97 baseline system has about seven thousands equivalence classes of HMM states. The NRDS estimate is obtained by splitting further (using lower thresholds) the baseline clustering tree to about 14 thousands classes. HMM state classes at Leaf nodes will be the "atoms" for which data sharing is carried out. At the end of splitting, $B(s)$ is determined as follows. For each leaf node s, all other leaf nodes are ordered by the divergence between underlying Gaussian distributions. The closest few nodes are selected as $B(s)$ so that the total occupancy is greater than a threshold. Gaussian parameters of new equivalence classes are set initially to that of their baseline parent nodes.

Subsequently sharing weights are optimized using the growth-transform [7]. Figure 1 depicts typical changes of an objective function and of data-sharing weights.

The NRDS model is used to rescore lattices generated by the baseline system. Results reported here are word-error-rate (WER) on the WS97 dev-test set, which consists of 2427 sentences and has about eighteen thousand words. A bigram language model is used in both systems. The baseline result is summarized on the first line of Table 1. The corresponding result for NRDS can be found on the line beginning with "NRDS-GT". We can see that NRDS model gives us 0.9% absolute WER reduction. The improvement is statistically significant.

As a test of importance of optimizing data sharing weights, models with two trivial sets of W are built and tested. In Table 1, the line marked with $W = 0$ is the result with $w(s,s) = 1$ and $w(u,s) = 0$ for all $u \neq s$. This is essentially a state-tied model with about 14 thousand equivalence classes of HMM states. The line with $W = 1$ means $w(u,s) = 1$ for all $u \in B(s)$ and 0 otherwise. This model corresponds to the highest degree of constraint on model parameters given $B(s)$ and 14 thousands equivalence classes. This suggests that the improvement is indeed due to better data sharing weights, not to the increased number of HMM equivalence classes.

	WER
Baseline	39.1
NRDS-GT	38.2
W=0	39.0
W=1	39.3

Table 1: Comparison of WER of NRDS vs. baseline system

5 Conclusions

In this paper we have developed a novel approach to estimating HMM parameters, namely non-reciprocal data sharing (NRDS). NRDS is obtained by generalizing parameter-tying, which can be regarded as reciprocal data sharing. A data sharing matrix is introduced to weigh the contributions from other HMM states when estimating model parameters of a HMM state.

NRDS can be viewed as smoothing the MLE of HMM parameters, where data sharing weights parameterize the degree of smoothing. We have shown that the data sharing matrix can be optimized by cross-validation. Under some assumptions, the objective function of cross-validation is a sum of rational functions of data-sharing weights. It is shown in this report that a sum of rational functions can be optimized efficiently by the growth-transform.

Our experiments show that NRDS reduces WER by 0.9% (absolute) on the LVCSR WS97 dev-test set. The reduction is statistically significant at confidence level 95%.

References

[1] L. E. Baum. An inequality and associated maximization techniques in statistical estimation for probabilistic functions of markov processes. *Inequalities*, 3:1–8, 1972.

[2] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, Series B*, 39(1):1–38, 1977.

[3] Frederick Jelinek ed. Research notes 14: 1997 lvcsr summer research workshop technical reports. Technical report, CLSP, The Johns Hopkins University, 1998.

[4] John J. Godfrey, Edward C. Holliman, and Jane McDaniel. Switchboard: Telephone speech corpus for research and development. In *Proc. of ICASSP*, volume I, pages 517–520, 1992.

[5] P. S. Gopalakrishnan, D. Kanevsky, Arthur Nadas, and David Nahamoo. An inequality for rational functions with applications to some statistical estimation problems. *IEEE Trans. on Information Theory*, 37(1):107–113, 1991.

[6] Frederick Jelinek and Robert Mercer. Interpolated estimation of Markov source parameters from sparse data. In E.S Gelsema and L. N. Kanal, editors, *Pattern Recognition in Practice*, North-Holland, Amsterdam, 1980.

[7] Xiaoqiang Luo. Growth transform of a sum of rational functions and its application in estimating HMM parameters. In *Proc. SST-98*, 1998.

[8] Xiaoqiang Luo and Frederick Jelinek. Nonreciprocal data sharing in estimating HMM parameters. Technical Report 32, CLSP, The Johns Hopkins University, 1998.

[9] Carey E. Priebe. Nonhomogeneity analysis using borrowed strength. *Journal of the American Statistical Association*, 91(436):1497–1503, December 1996.