interspeech 2008 incorporating SST 08

Maximum Accept and Reject (MARS) training of HMM-GMM speech recognition systems

Vivek Tyagi

IBM India Research Laboratory
New Delhi, India
vityagi1@in.ibm.com

Abstract

This paper describes a new discriminative HMM parameter estimation technique. It supplements the usual ML optimization function with the emission (accept) likelihood of the aligned state (phone) and the rejection likelihoods from the rest of the states (phones). Intuitively, this new optimization function takes into the account as to how well the other states are rejecting the current frame that has been aligned with a given state. This simple scheme, termed as Maximum Accept and Reject (MARS), implicitly brings in the discriminative information and hence performs better than the ML trained models. As is well known, maximum mutual information (MMI)[3, 4] training needs a language model (lattice), encoding all possible sentences[7, 9], that could occur in the test conditions. MMI training uses this language model (lattice) to identify the confusable segments of speech in the form of the so-called "denominator" state occupation statistics [7]. However, this implicitly ties the MMI trained acoustic model to a particular task-domain. MARS training does not face this constraint as it finds the confusable states at the frame level and hence does not use a language model (lattice) during training.

1. Introduction

The most popular hidden Markov model and Gaussian mixture model (HMM-GMM) training technique is the maximum likelihood (ML)[1, 2] technique where the optimization function is the likelihood of the feature vectors given the correct phone transcription.

$$F_{ML}(\lambda) = \log_\theta p(o_1^T|w_1^N) \tag{1}$$

where, θ is the set of all the HMM-GMM parameters, o_1^T is the sequence of T feature vectors corresponding to an utterance with the correct phonetic transcription w_1^N. This is the likelihood of the observations of training data given the correct transcription's composite HMM. It is well known that if the true distribution of the data lies in the space of the assumed family of the distributions, and if sufficient amount of training data is available, then the probability distribution function (pdf) parametrized by the ML estimates, converge to the true distribution of the data[1]. An additional reason for the widespread use of the ML estimation is that simple and efficient parameter estimation algorithm (E-M algorithm) exist for ML case[2].

In most of the speech recognition systems, the family of distributions modeling the acoustic observations is assumed to be a mixture of Gaussians. However, this assumption is not valid in practice. Therefore, using the model parameters that maximize the likelihood of the acoustic data conditioned on the correct phonetic transcription is not the best solution. It is also well know that increasing the likelihood does not necessarily result into the increased word or phone recognition accuracies. Therefore new optimization functions, such as, maximum mutual information (MMI)[3, 4, 5], minimum classification error (MCE) and minimum phone error (MPE)[9] were introduced which included certain discriminative criterion. Hybrid ANN-HMM[11] speech recognition systems also incorporated discriminative training through the use of the neural networks.

We consider the MMI objective function which is defined as:

$$\begin{aligned}\hat{\theta}_{MMI} &= \arg\max_\theta p_\theta(w_1^N|o_1^T) \\ &= \arg\max_\theta \frac{p_\theta(o_1^T|w_1^N)P(w_1^N)}{\sum_{w'} p_\theta(o_1^T|w_1'^N)P(w_1'^N)} \\ &\simeq \arg\max_\theta \frac{p_\theta(o_1^T|w_1^N)^\kappa P(w_1^N)^\kappa}{\sum_{w'} p_\theta(o_1^T|w_1'^N)^\kappa P(w_1'^N)^\kappa}\end{aligned} \tag{2}$$

where, $P(w_1^N)$ is the language model probability for sentence w_1^N and κ is an empirical factor to ensure that MMI training leads to a good test-set performance. The denominator term in (2) is the expected likelihood of the observables over all possible phone sequences $w_1'^N$. In practice, it is estimated using a lattice that in indicative of the test-case's language model. This implicitly couples the MMI training with a specific language model.

Going a step further, we can devise a discriminative objective function that does not consider all possible transcriptions as in (2) but rather considers frame level errors. We consider a maximum accept and reject (MARS) objective function as follows:

$$\hat{\theta}_{MARS} = \arg\max_\theta \prod_{t=1}^{T} \frac{P(q_t|q_{t-1})p(o_t|q_t)}{\prod_{q=1,q\neq q_t}^{Q} p(o_t|q)^\nu} \tag{3}$$

where q_t is the state aligned at time t, q_1^T is the correct state sequence as per the phonetic transcription and Q are the total number of states in the HMMs being trained. Qualitatively, MARS brings in discrimination at the frame level. Instead of just taking into account the accept (emission) likelihood $p(o_t|q_t)$, it also considers the reciprocal of the rejection likelihood from the rest of the states $\prod_{q=1,q\neq q_t}^{Q} p(o_t|q)^\nu$. In (3), ν is an empirical factor to control the influence of the accept and the reject likelihoods. We also note that unlike the MMI criterion, MARS is not coupled with a language model.

10.21437/Interspeech.2008-117

2. MARS Training

2.1. MARS objective function

Given the correct phonetic transcription of the training data a composite HMM can be made for the training utterance. In the following discussion, we will denote the hidden HMM state sequence by q_1^T where q_t is the state at time t. Following [2], the maximization in (2) may be expressed in terms of the hidden states of the HMM as follows;

$$\hat{\theta}_{MARS} = \arg\max_{\hat{\theta}} \prod_{t=1}^{T} \frac{P(q_t|q_{t-1}) p_{\hat{\theta}}(o_t|q_t)}{\prod_{q=1, q\neq q_t}^{Q} p_{\hat{\theta}}(o_t|q_t)^{\nu}} \quad (4)$$

Whereas, the ML objective function is,

$$\hat{\theta}_{ML} = \arg\max_{\hat{\theta}} \prod_{t=1}^{T} P(q_t|q_{t-1}) p_{\hat{\theta}}(o_t|q_t) \quad (5)$$

Comparing (4) and (5), we find that the only difference between the two objective functions is the extra term $\frac{1}{\prod_{q=1, q\neq q_t}^{Q} p_{\hat{\theta}}(o_t|q_t)^{\nu}}$. This is the rejection likelihood for the frame o_t when it has been aligned with the state q_t.

As the states q_t are hidden (their values are not known), direct maximization in (4) is not possible. Therefore following [2], we consider the expectation of (4) with respect to probability distribution of the states: $P_\theta(q_1^T|o_1^T)$. This leads to the following maximization,

$$\arg\max_{\hat{\theta}} E_{\theta, q_1^T|o_1^T}[\log \frac{\prod_{t=1}^{T} P_{\hat{\theta}}(q_t|q_{t-1}) p_{\hat{\theta}}(o_t|q_t)}{\prod_{q=1, q\neq q_t}^{Q} p_{\hat{\theta}}(o_t|q_t)^{\nu}}] \quad (6)$$

where θ represents the previous estimate of the model parameters, $\hat{\theta}$ represents the current estimates of the model parameters that will be estimated in the current iteration. $p_\theta()$ indicates that the pdf is parametrized by the parameters θ and $E_{\theta, q_1^T|o_1^T}$ denotes the expectation over q_1^T (the hidden state sequence) conditioned on o_1^T (the acoustic observables). From [2], this is equivalent to:

$$\arg\max_{\hat{\theta}} \sum_{q_1^T} P_\theta(q_1^T|o_1^T) \times \quad (7)$$

$$\sum_{t=1}^{T} \left\{ \log p_{\hat{\theta}}(o_t|q_t) - \sum_{q=1, q\neq q_t}^{Q} \log p_{\hat{\theta}}(o_t|q) + \log P_{\hat{\theta}}(q_t|q_{t-1}) \right\}$$

For the sake of brevity, we have not explicitly mentioned that the HMMs are made from the correct transcription. From [2], the solution of (7) is an iterative process with an (E)xpectation step followed by a (M)aximization. The E-step involves computation of the state posterior probability: [1] $p(q_t|o_1^T)$ and it may be computed quite efficiently with the Forward-Backward algorithm. The M-step involves choosing the parameters $\hat{\theta}$ to maximize (7). Using $\gamma_i^{ML}(t)$ to denote $p(q_t = i|o_1^T)$ and further making the simplifying assumption that each state (phone) is represented by a single one dimensional Gaussian, we get,

$$\arg\max_{\hat{\theta}} \sum_{i=1}^{Q} \sum_{t=1}^{T} \gamma_i^{ML}(t) \times \{\log p_{\hat{\theta}}(o_t|q_t = i) - \sum_{q=1, q\neq i}^{Q} \log p_{\hat{\theta}}(o_t|q) \} + \text{Transition Terms} \quad (8)$$

For the sake of simplicity, we will ignore the transition probability terms in (8). We also drop the explicit reference to the subscript θ in the pdfs. Differentiating (8) by the mean of the j^{th} state's Gaussian and setting it to zero, we get,

$$\sum_{t=1}^{T} \gamma_j^{ML}(t)(2(o_t - \mu_j)) - \sum_{t=1}^{T} \sum_{i=1, i\neq j}^{Q} \nu \gamma_i^{ML}(t)(2(o_t - \mu_j)) = 0 \quad (9)$$

Re-arranging the terms in (9) and denoting $\sum_{i=1, i\neq j}^{Q} \gamma_i^{ML}(t)$ as $\gamma_j^{Rej}(t)$ we get,

$$\mu_j^{MARS} = \frac{\sum_{t=1}^{T} \gamma_j^{ML}(t) o_t - \nu \times \sum_{t=1}^{T} \gamma_j^{Rej}(t) o_t}{\sum_{t=1}^{T} \gamma_j^{ML}(t) - \nu \times \sum_{t=1}^{T} \gamma_j^{Rej}(t)} \quad (10)$$

Let us further consider the quantity: $\gamma_j^{Rej}(t) = \sum_{i=1, i\neq j}^{Q} \gamma_i^{ML}(t)$. Typically for a given frame t, $\gamma_k^{ML}(t) = 1$ for only a particular state k and is zero for the rest of the states. This results in $\gamma_k^{Rej}(t) = 0$ and $\gamma_i^{Rej}(t) = 1, i \neq k$. Therefore this particular frame contributes as a positive example for the accept (emission) density of the k^{th} state and as a negative example for the rest of the states. In words, once a particular frame 'o_t' has been assigned to a particular state k by the virtue of $(\gamma_k^{ML}(t) = 1)$[2], MARS training moves the mean of the remaining states away from the 'o_t'. Now let us compare μ_j^{MARS} with μ_j^{ML} and μ_j^{MMI} given below,

$$\mu_j^{ML} = \frac{\sum_{t=1}^{T} \gamma_j^{ML}(t) o_t}{\sum_{t=1}^{T} \gamma_j^{ML}(t)} \quad (11)$$

$$\mu_j^{MMI} = \frac{\sum_{t=1}^{T} \gamma_j^{ML}(t) o_t - \sum_{t=1}^{T} \gamma_j^{Den.}(t) o_t + D\mu_j^{ML}}{\sum_{t=1}^{T} \gamma_j^{ML}(t) - \sum_{t=1}^{T} \gamma_j^{Den.}(t) + D}$$

where the constant D is set on a per-Gaussian level according to certain rules[5, 7]. $\gamma_j^{ML}(t)$ and $\gamma_j^{Den.}(t)$ are obtained by applying the Forward-Backward algorithm on the observables o_1^T using the correct phonetic transcription and the recognition lattice respectively[7, 8]. As can be noted in (10) and (11), both MARS and MMI are correcting the "state occupation counts" and their moments through $\gamma_j^{Rej}(t)$ and $\gamma_j^{Den.}(t)$ respectively. However there is a major difference in the way the "state occupation counts" are computed for the two cases.

- In the case of MMI, $\gamma_j^{Den.}(t)$ is computed completely independent of the $\gamma_j^{ML}(t)$[3].
- Whereas in the case of the MARS, $\gamma_j^{Rej}(t)$ is computed in a coupled manner with the $\gamma_i^{ML}(t)$.

The following is a pseudo code for the MARS estimation of the mean of the j^{th} state's Gaussian density.

[1] also called state occupation counts

[2] Or more generally $\gamma_k^{ML}(t)$ being the highest amongst all the states

[3] i.e. the alignment with respect to the correct phonetic transcription

1: **for all** $t \in (1,T), i \in (1,Q)$ **do**
2: Compute $\gamma_i^{ML}(t)$ using the correct phonetic transcription.
3: Initialize: $\gamma_i^{Rej}(t) = 0$
4: **end for**
5: **for all** $t \in (1,T), i \in (1,Q), i \neq j$ **do**
6: **if** $\gamma_i^{ML}(t) = 1$ and $p(o_t|q_i) \leq p(o_t|q_j)$ **then**
7: $\gamma_j^{Rej}(t) = \gamma_j^{Rej}(t) + \gamma_i^{ML}(t)$
8: **end if**
9: **end for**
10: $\mu_j^{MARS} = \frac{\sum_{t=1}^{T} \gamma_j^{ML}(t) o_t - \nu \times \gamma_j^{Rej}(t) o_t}{\sum_{t=1}^{T} \gamma_j^{ML}(t) - \nu \times \sum_{t=1}^{T} \gamma_j^{Rej}(t)}$

For a given frame 't', if $\gamma_i^{ML}(t) = 1$, then this implies that the i^{th} state is well aligned with the observable o_t given the correct phonetic transcription. Now consider a state $j, j \neq i$ such that $p(o_t|q_j) >= p(o_t|q_i)$. This corresponds to the line 6 in the above pseudo-code and it forms an important step of the MARS training technique. This condition indicates that in the absence of the correct phonetic transcription[4], this frame may be aligned with the incorrect state j. This would lead to a frame level classification error. Therefore, for all such states we set $\gamma_j^{Rej}(t) = \gamma_i^{ML}(t)$. In [8], the authors have also proposed a "frame discrimination training of HMMs". However, there are two key differences between the MARS and the technique in [8].

- In [8], the authors have used the MMI criterion with the denominator HMM allowing transitions between all possible states in the system. Therefore, for any given frame o_t, the same state q_i may be aligned by the numerator HMM and the denominator HMM as in (2). By its definition (3), this situation does not arise in MARS.
- MARS explicitly tracks the frame level errors as in the line: 6 of the pseudo code above whereas the technique in [8] does not compare the emission likelihoods between the numerator aligned state and the denominator aligned state.

2.2. Complete MARS re-estimation formulas

Now, let us consider the general case when the accept (emission) density is modeled by a Gaussian mixture model (GMM) with M components and diagonal covariance matrices. We denote the posterior probability of being in the state 'j' and the mixture component being 'l' as $\gamma_{j,l}^{ML}(t)$. Let $p_{i,l}(o_t)$ be the likelihood of o_t being emitted i^{th} state's l^{th} mixture component. Then $\gamma_{j,l}^{Rej.}(t) = \sum_{i=1,ij}^{Q} \gamma_i^{ML}(t) \frac{p_{i,l}(o_t)}{p_i(o_t)}$. The parameter re-estimation formulas for the mean ($\mu_{j,l}^{MARS}$), variance($\sigma_{j,l}^{MARS}$) and the l^{th} Gaussian's weight ($c_{j,l}$) are,

$$\mu_{j,l}^{MARS} = \frac{\sum_{t=1}^{T} \gamma_{j,l}^{ML}(t) o_t - \nu \times \sum_{t=1}^{T} \gamma_{j,l}^{Rej}(t) o_t}{\sum_{t=1}^{T} \gamma_{j,l}^{ML}(t) - \nu \times \sum_{t=1}^{T} \gamma_{j,l}^{Rej}(t)} \quad (12)$$

$$\sigma_{j,l}^{MARS} = \frac{\sum_{t=1}^{T} \gamma_{j,l}^{ML}(t) o_t^2 - \nu \times \gamma_{j,l}^{Rej}(t) o_t^2}{\sum_{t=1}^{T} \gamma_{j,l}^{ML}(t) - \nu \times \sum_{t=1}^{T} \gamma_{j,l}^{Rej}(t)} - (\mu_{j,l}^{MARS})^2$$

$$c_{j,l}^{MARS} = \frac{\sum_{t=1}^{T} \gamma_{j,l}^{ML}(t) - \nu \times \gamma_{j,l}^{Rej}(t)}{\sum_{t=1}^{T} \gamma_j^{ML}(t) - \nu \times \gamma_j^{Rej}(t)}$$

2.3. Setting the factor ν

In the MARS training, the rejection state occupation counts ($\sum_{t=1}^{T} \gamma_j^{Rej}(t)$) are usually much greater than the usual state occupation counts ($\sum_{t=1}^{T} \gamma_j^{ML}(t)$. Therefore, the factor ν in (9) is set as: $\nu = 0.35 \times \frac{\sum_{t=1}^{T} gamma_j^{ML}(t)}{\sum_{t=1}^{T} gamma_j^{Rej}(t)}$ which effectively assigns 0.35 weight to the "normalized rejection" moment for each state 'j'.

[4] as in the test conditions

3. Experiments

We have implemented and tested the MARS training procedure for the context-independent phoneme recognition on the TIMIT corpus. The 61 TIMIT phones are first mapped to 48 phones for training and are finally folded to 39 phones for testing. Each mono-phone is modeled by a 3 state left-to-right context-independent HMM with no skip states. The Accept (emission) density in each state is modeled by mixture of 11 component diagonal covariance Gaussians[5] for both the ML and MARS system. We use a bi-gram language model (trained on the train-set) for both the ML and MARS testing[6]. The phoneme recognition accuracies on the TIMIT *core-test set* using various models are provided in Table. 1. The MARS training was initialized with 11 component ML model and the number in the brackets indicates the number of MARS iterations. As can be seen from the Table. 1, the best accuracy of 69.1 is achieved just after the first iteration of MARS. The ML and the MARS training and recognition is performed using IBM-IRL's HMM training toolkit (IrlTK).

Table 1: Phoneme recognition accuracy (including the substitution, deletion and insertion errors) using the ML and MARS models.

System(itr)	Accuracy
ML	67.6
MARS(1)	69.1
MARS(2)	68.5
MARS(3)	68.6

In [6], the authors have reported the phoneme recognition accuracies on the TIMIT corpus using the ML and MMI training techniques. The output density is modeled by 16 component diagonal covariance Gaussians and a bi-gram language model was used. As reported by the authors the MMI performance peaked after about 8 iterations. For comparison, we are reporting the results directly from their paper[6]

Table 2: Phoneme recognition accuracy using MMI[6]

System(itr)	Accuracy
ML[6]	66.1
MMI[6]	67.5

From the Tables.1 and 2 we find that the MARS recognition accuracy is comparable to that of MMI. However, we are unable explain the difference between the ML accuracy obtained through IrlTK (Table.1) and the ML accuracy in [6] even though the training and the testing conditions are similar.

In [10], the authors have used Extended Baum-Welch (EBW) transformations in the context of the HMMs for the recognition

[5] The ML phone recognition accuracy saturated around 11 component Gaussian mixture

[6] The insertion penalty (-2) and the LM weight (4) was tuned for the ML case and the same values were used for testing the MARS model

of 7 broad phonetic classes on the TIMIT corpus. The 7 broad phonetic classes (BPC) are (Vowels/Semi-Vowels, Nasal/Flaps, Strong Fricatives, Weak Fricatives, Stops, Closures, Silence). All the 61 labels, except the glottal stop 'q', are mapped into these 7 BPC. We further compared the MARS accuracy with the EBW gradient metric[10] accuracy on the BPC recognition task and the results are provided in Table. 3. We note that the MARS results are comparable to the EBW gradient metric based system.

Table 3: BPC recognition accuracy (including the substitution, deletion and insertion errors) on the *TIMIT core-test set* using the ML, EBW and MARS models.

System(itr)	Accuracy
ML[10]	80.5
EBW-F[10]	80.1
EBW-F Norm[10]	81.1
ML(IrlTK)	81.5
MARS(1)	82.5

4. Conclusions and Future work

In this paper we have formulated a new discriminative HMM-GMM parameter estimation technique that takes frame level phoneme classification errors into account while learning the parameters of the phoneme distributions. Unlike MMI/MPE, MARS training does not require a language model or recognition lattice to extract all the confusable segments. This can perhaps be advantageous in the real-life systems where the test-case language models can change over time. Our initial context-independent experiments indicate that the MARS technique is comparable to other discriminative techniques such as the MMI and the EBW. In the future, we would like to expand this work to the more complex tied, context-dependent phoneme models and the large vocabulary recognition tasks.

5. References

[1] A. Nadas, "A decision theoretic formulation of a training problem in speech recognition, and a comparison of training by conditional versus un-conditional maximum likleihood", IEEE Trans. Acoustics, Speech and Sig. Proc., pp814-817, Aug 1983.

[2] A. P. Demspter, N. M. Laird and D. B. Rubin, "Maximum likelihood estimation from incomplete data", Journal of the Royal Statistical Society (B), vol. 39, no. 1, pp1-38, 1979.

[3] L. R. Bahl, P. F. Brown, P.V. de Souza and R. L. Mercer, "Maximum mutual information estimation of the hidden Markov parameters for speech recognition", In Proc. of IEEE ICASSP, pp49-52, 1986.

[4] P.S. Gopalakrishnan, D. Kanevsky, A. Nadas and D. Nahamoo, "An inequality for rational functions with applications to some statistical estimation problems", IEEE Tran. on Information Theory, Vol. 37, No.1, pp 107-113, Jan 1991.

[5] Y. Normandin, "Optimal splitting of HMM Gaussian mixture components with MMIE training ", In the Proc. of IEEE ICASSP, pp-449-452, 1995.

[6] S. Kapadia, V. Valtchev and S.J. Young, "MMI Training for continuous phoneme recognition on the TIMIT database", In the Proc. of IEEE ICASSP, pp 491-494, 1993.

[7] V. Valtchev, J.J. Odell, P.C. Woodland and S.J. Young, "MMIE training of large vocabulary recognition systems", Speech Communication, Vol. 22, pp 303-314, 1997.

[8] D. Povey and P.C. Woodland, "Frame discrimination training for HMMs for large vocabulary speech recognition", In Proc. of IEEE ICASSP, Vol. 1, pp 15-19, March 1999.

[9] D. Povey and P.C. Woodland, "Minimum phone error and I-smoothing for improved discriminative training", In the Proc. of IEEE ICASSP, 2002.

[10] T.N. Sainath, D. Kanevksy and B. Ramabhadran, "Broad phonetic class recognition in a hidden Markov model framework using extended baum-welch transformations", In the Proc. of IEEE ASRU, 2007.

[11] N. Morgan and H. Bourlard, "Continuous speech recognition: An introduction to the Hybrid HMM/Connectionist Approach," IEEE Signal Processing Magazine, vol.12, no.3, pp.25-42, May 1995.