Discriminative Acoustic Modeling for Speech Recognition
- Maximum Entropy Direct Models -

Yueng-Tien, Lo
g96470198@csie.ntnu.edu.tw
Speech Lab, CSIE
National Taiwan Normal University

H-K. Kuo and Y. Gao, “Maximum entropy direct models for speech recognition,” IEEE Transactions Audio Speech and Language Processing,2006.

outline

- Introduction
- Direct Models
- Maximum Entropy Markov Models (MEMM)
- MEMM for Speech Recognition
- Experimental Results
- Discussion
- Conclusion

introduction

- Traditional statistical modes for speech recognition have mostly been based on a Bayesian framework using generative models such as Hidden Markov models (HMMs).

$$\hat{S} = \arg\max_{S} p(S \mid O).$$

•When solving with the HMM, $\hat{S}$ is obtained by maximizing the joint probability

$$\hat{S} = \arg\max_{S} \frac{p(O \mid S)\, p(S)}{p(O)} = \arg\max_{S} p(O \mid S)\, p(S)$$

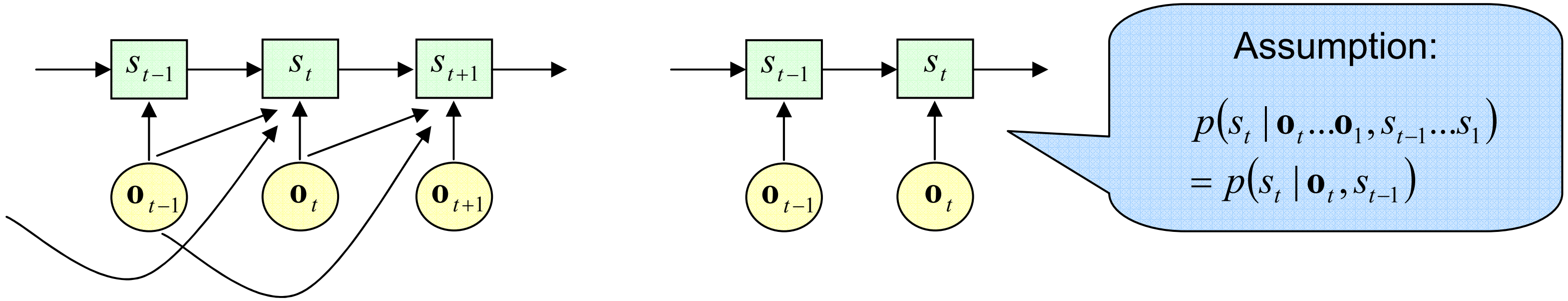
Direct Models

- Direct modeling attempts to model the posterior probability $P(S \mid O)$ directly

- There are many potential advantages as well as challenges for direct modeling
  - The direct model can potentially make decoding simpler
  - The direct model allows for the potential combination of multiple sources of data in a unified fashion
    - Asynchronous and overlapping features can be incorporated formally
    - It will be possible to take advantage of supra-segmental features like prosodic features, acoustic phonetic features, speaker style, rate of speech, channel differences
  - However, joint estimation would require a large amount of parallel speech and text data (a challenge for data collection)

Direct Models

- The relationship between observations and states is reversed
  - Separate transition and observation probabilities are replaced with one function $p(s_t \mid \mathbf{o}_t, s_{t-1})$
  - Directly modeling $p(s_t \mid \mathbf{o}_t, s_{t-1})$ makes direct computation of $P(\mathbf{S} \mid \mathbf{O})$ possible
- The model can also be conditioned flexibly on a variety of contextual features
  - Any computable property of the observation sequence can be used as a feature
  - The number of features at each time frame need not be the same



Assumption:

$$p(s_t \mid \mathbf{o}_t \ldots \mathbf{o}_1, s_{t-1} \ldots s_1) = p(s_t \mid \mathbf{o}_t, s_{t-1})$$

Maximum Entropy Markov Models

- Recently, *McCallum et al.* (ICML 2000) modeled sequential processes using a direct model similar to the HMM in graphical structure and used exponential models for transition-observation probabilities
  - Called Maximum Entropy Markov Model (MEMM)

- Maximum Entropy modeling is used to model the conditional distributions
  - ME modeling is based on the principle of avoiding unnecessary assumptions
  - The principle states that **the modeled probability distribution should be consistent with the given collection of facts about itself and otherwise be as uniform as possible**

Maximum Entropy Markov Models

- The mathematical interpretation of this principle results in a constrained optimization problem
  - Maximize the entropy of a conditional distribution $p(s_t \mid \mathbf{o}_t, s_{t-1})$, subject to given constraints
  - Constraints represent the known facts about the model from statistics of the training data

Definition 1:

$$f_{\langle b,s \rangle}(\bar{c}_t, s_t) = \begin{cases} \alpha > 0, & \text{if } b(\bar{c}_t) = 1 \text{ and } s_t = s \\ 0, & \text{otherwise} \end{cases}$$

$$\text{in MEMM, } \bar{c}_t = \{\mathbf{o}_t, s_{t-1}\}$$

Definition 2:

f_i is said to be activated by a given pair $\langle \bar{c}_t, s_t \rangle$, if $f_i(\bar{c}_t, s_t) \neq 0$

Maximum Entropy Markov Models

- These definitions allow us to introduce the constraints of the model

$$\forall f_i : E_{f_i} = \widetilde{E}_{f_i}$$

$$\widetilde{E}_{f_i} = \sum_{\bar{c}' \in C', s \in V} \widetilde{p}(\bar{c}', s) f_i(\bar{c}', s) = \frac{1}{N} \sum_{j=1}^{N} f_i(c_j, s_j)$$

- The expected value of f_i with respect to the model $p(s \mid \bar{c})$ is

$$E_{f_i} = \sum_{\bar{c}' \in C', s \in V} p(s \mid \bar{c}') \widetilde{p}(\bar{c}') f_i(\bar{c}', s)$$

- Using Lagrange multipliers for constrained optimization, the desired probability distribution is given by the maximum of the function

$$\Lambda(p(s \mid \bar{c}), \lambda) = H(p(s \mid \bar{c})) + \sum_i \left(\lambda_i \cdot \left(E_{f_i} - \widetilde{E}_{f_i} \right) \right)$$

$$H(p(s \mid \bar{c})) = - \sum_{\bar{c}' \in C', s \in V} \widetilde{p}(\bar{c}') p(s \mid \bar{c}') \log p(s \mid \bar{c}')$$

Maximum Entropy Markov Models

- Finally, the solution of objective function is given by an exponential model

$$p_\lambda(s \mid \bar{c}) = \frac{\exp\left(\sum_i \left(\lambda_i \cdot f_i(\bar{c}, s)\right)\right)}{Z_\lambda(\bar{c})}$$

where the summation is over all features of the model and $Z_\lambda(\bar{c})$ is a normalization factor given by

$$Z_\lambda(\bar{c}) = \sum_{s' \in V} \exp\left(\sum_j \left(\lambda_j \cdot f_j(\bar{c}, s')\right)\right)$$

Direct models for Speech Recognition

- The typical vectors are mel-frequency cepstra coefficients (MFCCs), with an option of applying LDA rotation
- The acoustic feature vector for every frame is quantized into a list of Gaussian identities (IDs) ranked in the order of their Gaussian likelihoods. $P(O_t \mid g_{ID})$
- The size of the list can be as large as the number of Gaussians used in the system, which is about 45k in our experiment.

- For simplification, only the top N Gaussian IDs are considered .(N =10)

- Hence, an observation $\overline{o}$ is a vector of size N, where each element is an integer with a value range of M(M=45K or 2048)

$$\overline{o} = g_1 g_2 \cdots g_N$$

- The states are defined as context-independent subphones.We use three subphone states for each of 52 phones, resulting in a state space V of size 156.

Feature for Phoneme Recognition

•The primary role of features is to trace correlations between various characteristics of the contexts and corresponding states.
Here two types of features are used:

•Observation Features: These features express dependencies between target states and appearance of particular Gaussian IDs in observation vectors:

$$f_{\langle \hat{g},s \rangle}(\overline{c}_t, s_t) = \begin{cases} \alpha > 0, & \text{if } \hat{g} \in \overline{o}_t \text{ and } s_t = s \\ 0, & \text{otherwise} \end{cases}$$

$$\alpha = \frac{2}{(1+R)}$$

Where R is the rank of the Gaussian. Notice that for the best matching Gaussian R=1 and α =1.0

Feature for Phoneme Recognition

- Transition Features:

  In most sequential processes, transition probabilities are not uniform
  This information can be used by adding special features.

$$f_{\langle \hat{s},s \rangle}\left(\bar{c}_t, s_t\right) = \begin{cases} 1, & \text{if } s_{t-1} = \hat{s} \text{ and } s_t = s \\ 0, & \text{otherwise} \end{cases}$$

- An important issue here is feature selection.

Decoding

The decoding algorithm used to find the optimal state sequence from a given observation sequence can be formulated as a Viterbi decoding procedure

$$\begin{aligned}
& p_\lambda(s_1 \ldots s_{t+1} \mid o_1 \ldots o_T) \\
&= p_\lambda(s_{t+1} \mid s_1 \ldots s_t, o_1 \ldots o_T)\, p_\lambda(s_1 \ldots s_t \mid o_1 \ldots o_T) \\
&= p_\lambda(s_{t+1} \mid s_t, o_{t+1})\, p_\lambda(s_1 \ldots s_t \mid o_1 \ldots o_T) \\
&= \prod_{j=1}^{t+1} p_\lambda(s_j \mid s_{j-1}, o_j)
\end{aligned}$$

Use dynamic programming to compute the state sequence with the highest probability

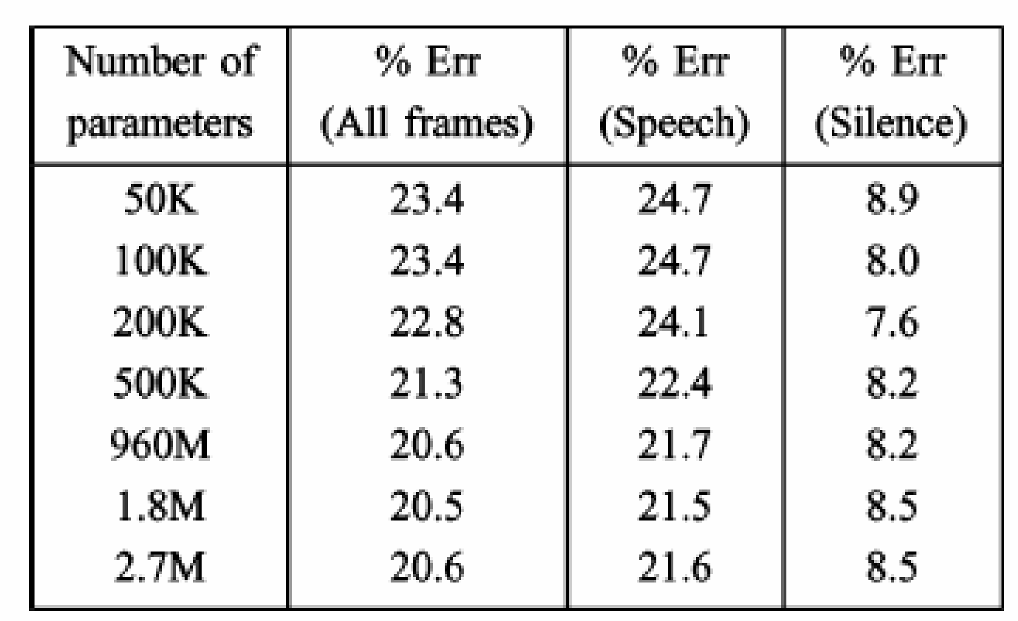
$$\hat{S} = \arg\max_{S} p_\lambda(s_1 \ldots s_T \mid o_1 \ldots o_T)$$

Experimental Setup

- A subset of the DARPA Communicator data, collected within the domain of air travel reservation.(380 h of speech)
- 240 K utterances (around 200 h) of speech were used for training the MEMM
- all the available training speech data were used to train the HMM.

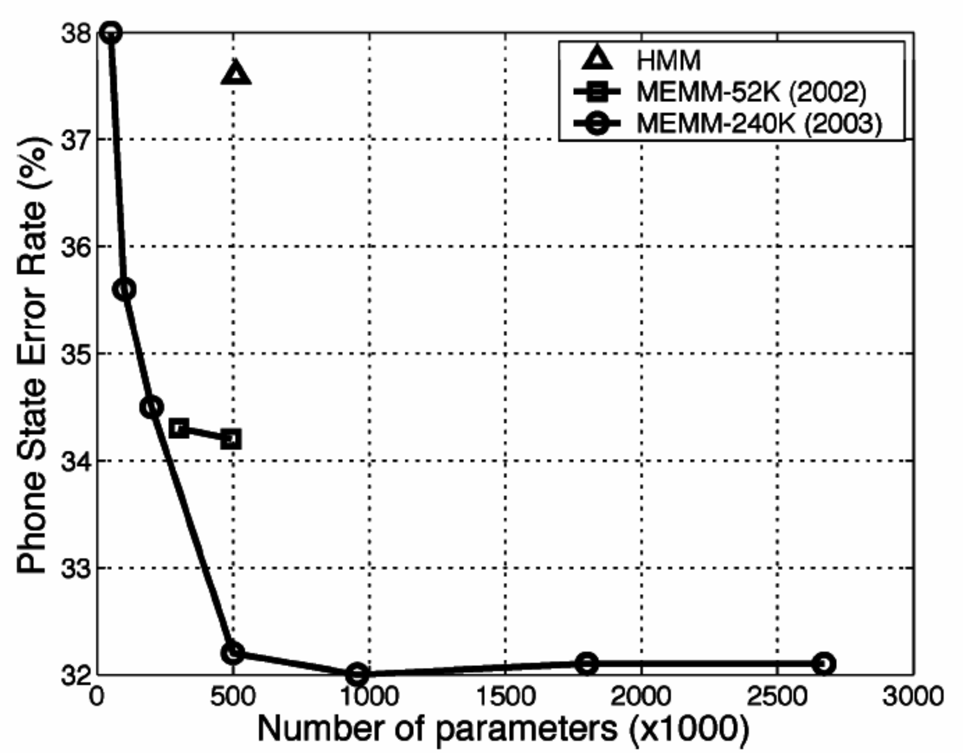
Experimental Results

- Frame classification error rate of the MEMM Acoustic Model with Different Complexity.
- Notice that there is a significant decrease in error rate when the model complexity increases from 200k parameters to 500k, and that the performance levels off after about 1 million parameters.
- Focusing on the speech frames, the model complexity increases from 200k to 500k.



Number of parameters	% Err (All frames)	% Err (Speech)	% Err (Silence)
50K	23.4	24.7	8.9
100K	23.4	24.7	8.0
200K	22.8	24.1	7.6
500K	21.3	22.4	8.2
960M	20.6	21.7	8.2
1.8M	20.5	21.5	8.5
2.7M	20.6	21.6	8.5

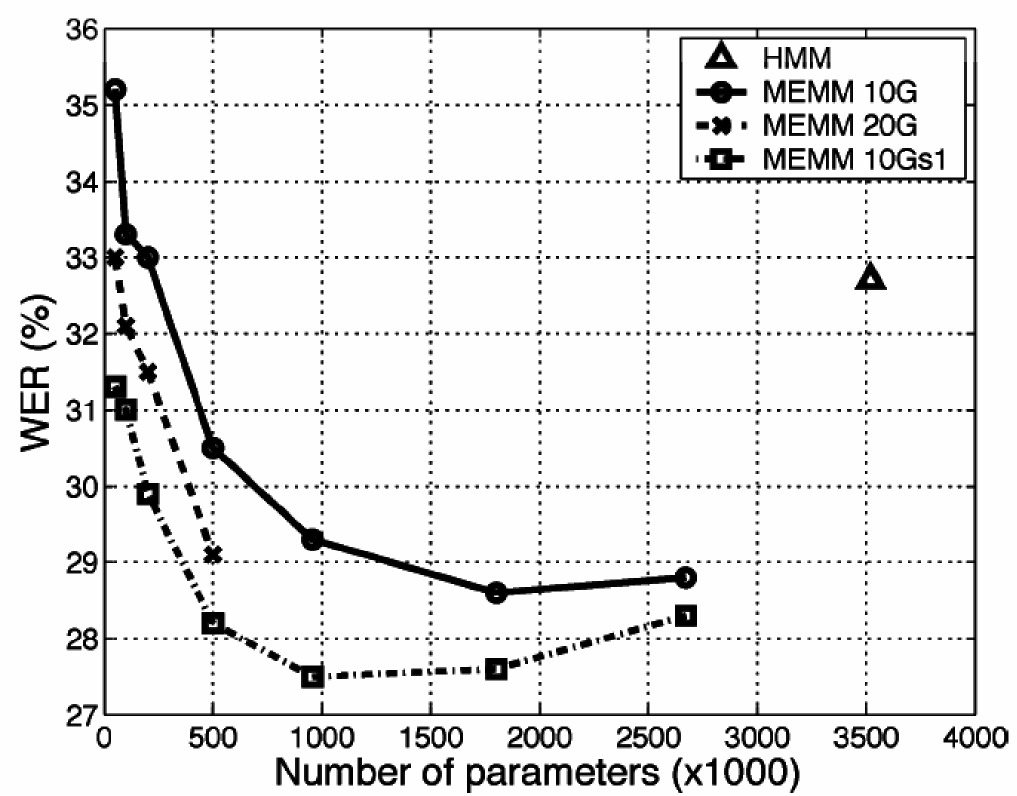
Phone Recognition Experiments



- Phone decoding results showing phone state recognition error rate for nonsilent frames. For the HMM, the error rate is 37.6%. MEMMs trained on more data outperforms those trained on less data

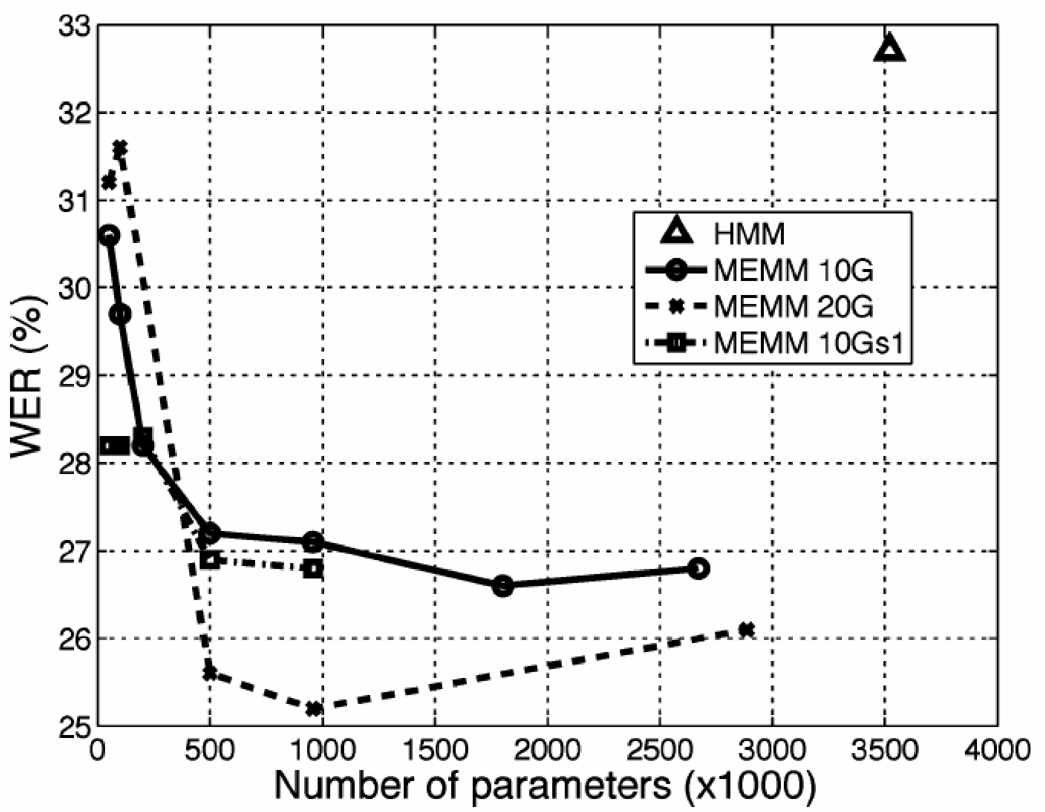
Word Recognition Through N-Best Rescoring(1/2)

- Compare the performance of the HMM and the new MEMMs as stand-alone acoustic models for word recognition through N-best hypotheses rescoring.
- The word error rate of different acoustic models when used to rescore the N-best hypotheses without incorporating the language model probabilities.



Word Recognition Through N-Best Rescoring(2/2)

- WER of stand-alone MEMM acoustic models, after many additional iterations of the IIS algorithm.



Combining with Language Model(1/3)

- A simple way to combine the language model would be to combine the language model with the acoustic model, with a weighting factor,as usually done in HMM speech recognition systems.

$$P(W \mid O) = \frac{P(O \mid W)P(W)}{P(O)} \propto p(O \mid W)^{\gamma} P(W)$$

- For the HMM, we typically use a log probability score such as

$$Q_{HMM+LM}(O_i, W_{ij}) = \gamma \cdot Q_{HMM}(Q_i, W_{ij}) + Q_{LM}(O_i, W_{ij})$$

$$Q_{HMM}(O_i, W_{ij}) = \log p_{HMM}(O_i \mid W_{ij})$$

$$Q_{LM}(O_i, W_{ij}) = \log P_{LM}(W_{ij})$$

For MEMM, a *similar combination*

$$Q_{MEMM+LM}(O_i, W_{ij}) = \gamma \cdot Q_{MEMM}(Q_i, W_{ij}) + Q_{LM}(O_i, W_{ij})$$

where

$$Q_{MEMM}(O_i, W_{ij}) = \log P_{MEMM}(W_{ij} \mid O_i)$$

Combining with Language Model(2/3)

- This naive combination method did not lead to good results, compared with the HMM recognizer.
- One reason may be that the direct MEMM probability is already a posterior probability model.

- Heuristically one can v.iew the MEMM, HMM, and language model as three different scores that can be combined to determine the best hypothesis in the N-best list

$$Q_{MEMMsc}(O_i, W_{ij}) = \beta_i \cdot [Q_{MEMM}(O_i, W_{ik)} - \min_k Q_{MEMM}(O_i, W_{ik})] + \min_k Q_{HMM}(O_i, W_{ik})$$

Where the utterance specific scaling factor is defined as

$$\beta_i = \frac{\max_k Q_{HMM}(O_i, W_{ik}) - \min_k Q_{HMM}(O_i, W_{ik})}{\max_k Q_{MEMM}(O_i, W_{ik}) - \min_k Q_{MEMM}(O_i, W_{ik})}$$

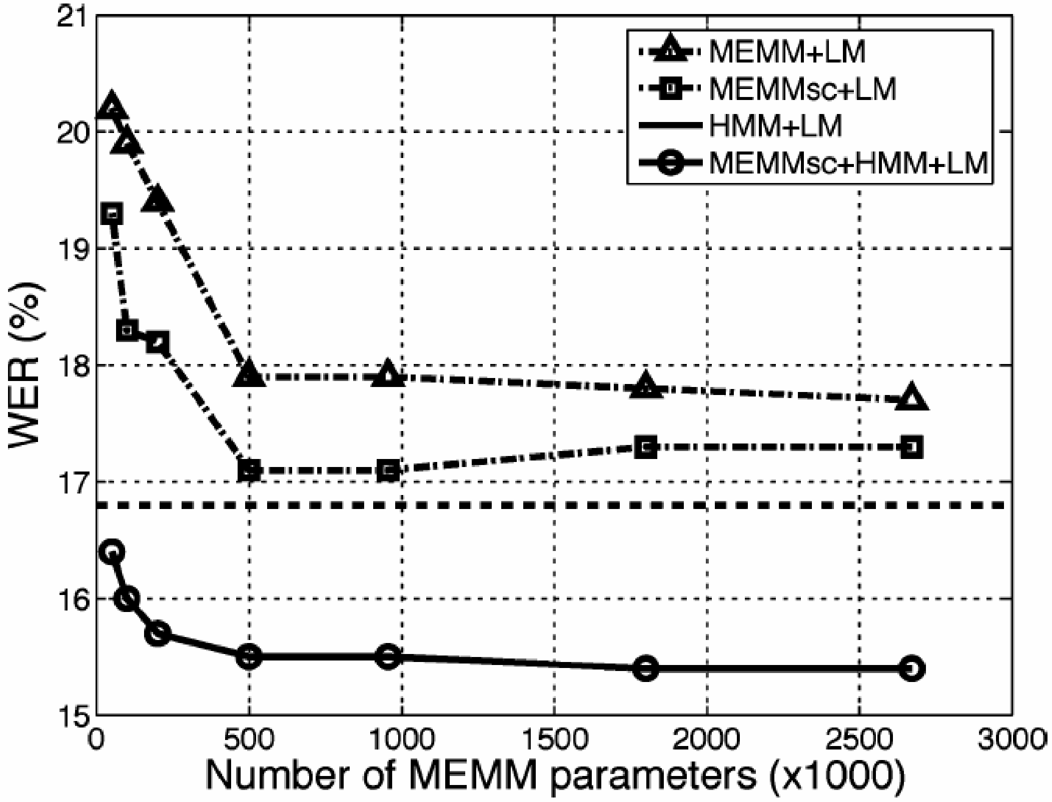
Combining with Language Model(3/3)

- Finally , we can define a new acoustic model score as the average of the scaled MEMM score and the HMM score

$$Q_{AC}(O_i,W_{ij})=\frac{1}{2}Q_{MEMMsc}(O_i,W_{ij})+\frac{1}{2}Q_{HMM}(O_i,W_{ij})$$

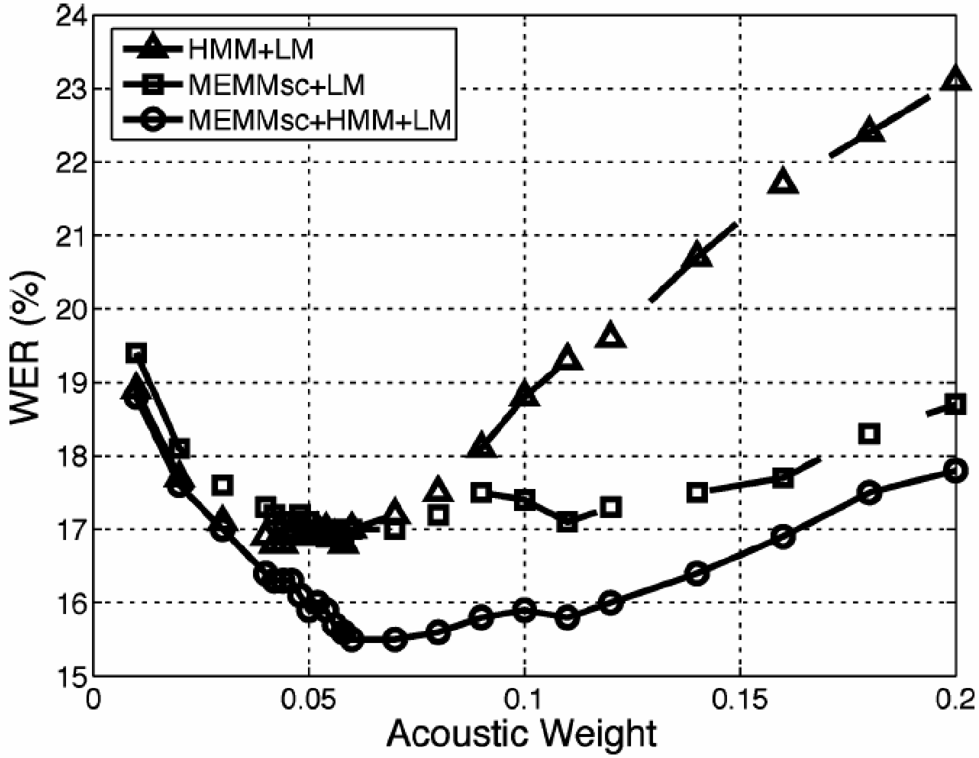
- And the combined score of MEMM, HMM, and LM scores to be

$$Q_{MEMMsc+HMM+LM}(O_i,W_{ij})$$
$$=\gamma\cdot Q_{AC}(O_i,W_{ij})+Q_{LM}(O_i,W_{ij})$$



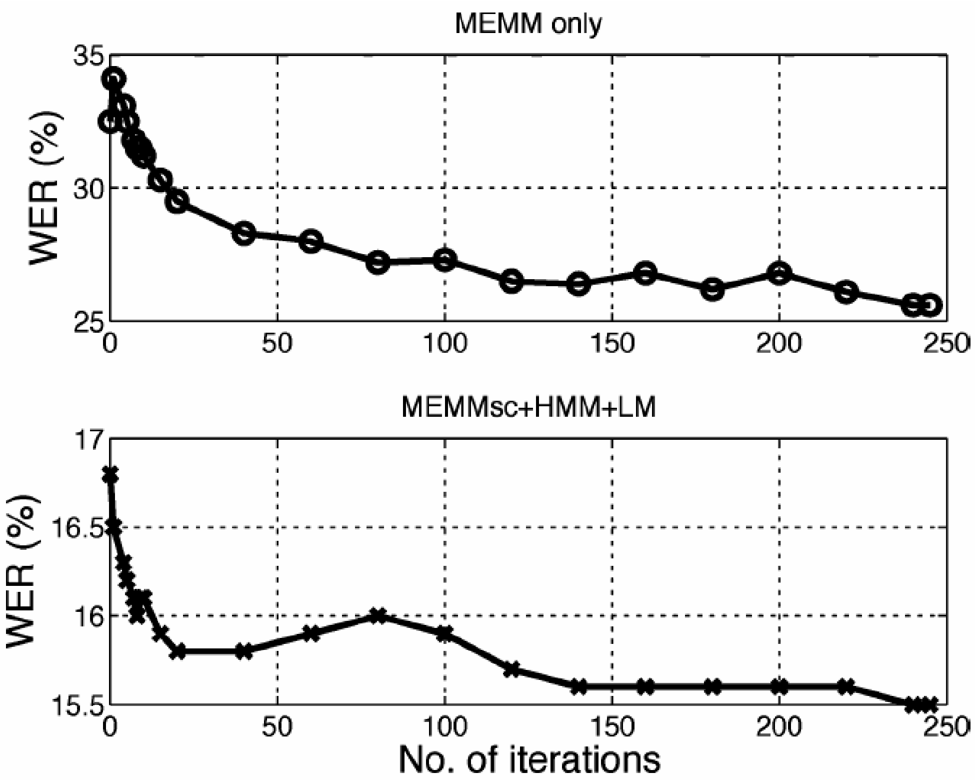
Combine MEMM with HMM

- Comparing the HMM and MEMM, the figure shows that the best performance (minimum WER) of the MEMM (square points) may be comparable or slightly worse than that of the HMM (triangle points).
- For the combined MEMM and HMM, the range of acoustic weights having low WER is significantly enlarged, compared to the results for the HMM alone (triangle).



Convergence behavior of IIS

- Approximately 250 iterations were necessary to obtain the model with optimal performance in WER.



Discussion

- Why the maximum entropy direct model using simple features in this paper outperforms the HMM ?

1. the feature space mapping (from acoustic feature vectors to ranked Gaussians) imparts a degree of robustness
2. the maximum entropy model is also the one that maximizes the likelihood of the training data

$$\hat{\lambda} = \arg\max \sum_{k=1}^{N} \log p_{\lambda}(s_{\kappa} \mid \bar{c}_{k})$$

- the MEMM incorporates a better model of the phonotactics through the transition features than the HMM. Some experimental results (not shown) indicate that the transition features can sometimes be helpful in the MEMM, whereas it is well known that they are not essential in the HMM.

Discussion

- The MEMM may suffer from a problem known as the “label bias” problem, which occurs for states whose outgoing transitions have low entropy, such that the observation features are effectively ignored.
  - This problem can be avoided by using another type of model known as conditional random fields (CRFs)

Conclusion

- In this paper, we challenge one of the commonly held assumptions that speech recognition must use generative models like HMMs and Bayes' rule in the maximum *a posteriori* decision rule to find the optimal word sequence

- The MEMM outperforms the HMM in terms of phone decoding and word decoding when used as an acoustic model, despite being trained on much less training data and having only the top ten ranked Gaussians as features