Maximum Entropy Modeling of Acoustic and Linguistic Features

Chuang-Hua Chueh and Jen-Tzung Chien

Department of Computer Science and Information Engineering
National Cheng Kung University, Tainan, Taiwan, ROC
{chchueh, chien}@chien.csie.ncku.edu.tw

ABSTRACT

Traditionally, speech recognition system is established assuming that acoustic and linguistic information sources are independent. Parameters of hidden Markov model and *n*-gram are estimated individually and then plugged in a maximum *a posteriori* classification rule. However, acoustic and linguistic features are correlated in essence. Modeling performance is limited accordingly. This study aims to relax the independence assumption and achieve sophisticated acoustic and linguistic modeling for speech recognition. We propose an integrated approach based on *maximum entropy (ME)* principle where acoustic and linguistic features are optimally merged in a unified framework. The correlations between acoustic and linguistic features are explored and properly represented in the integrated models. Due to the flexibility of ME model, we can further combine other high-level linguistic features. In the experiments, we carry out the proposed methods for broadcast news transcription using MATBN database. We obtain significant improvement compared to conventional speech recognition system using individual maximum likelihood training.

1. INTRODUCTION

Automatic speech recognition has been increasingly important in many human-machine interaction systems. How to build a desirable classification procedure is critical to assure system performance. Following Bayesian decision theory, speech recognition endeavors to find the most likely word sequence $\hat{W}$ through maximizing *a posteriori* (MAP) probability given an observed speech sentence X

$$\hat{W} = \arg\max_{W} p(W|X) = \arg\max_{W} p_{\Lambda}(X|W) p_{\Gamma}(W) . \quad (1)$$

In (1), $p_{\Lambda}(X|W)$ represents acoustic likelihood of matching signal X with hidden Markov models (HMM's) for W. The prior probability $p_{\Gamma}(W)$ serves as language model characterizing the linguistic regularities in natural language. *N*-gram model is popular to explore local lexical characteristics from text documents. Undoubtedly, the estimation of HMM's Λ and *n*-grams Γ plays an important role in speech recognition system. In the literature, maximum likelihood (ML) criterion is widely applied for parameter estimation. The estimated parameters can attain the largest likelihood score using training data. However, higher likelihood does not guarantee better classification. To improve classification performance, it is beneficial to directly enhance model discriminability. The minimum classification error (MCE) and maximum mutual information (MMI) criteria were proposed for discriminative modeling of acoustic [1][9][14] as well as linguistic features [2][11]. Under the assumption of independence between acoustic and linguistic events, parameters of HMM and *n*-gram were optimized individually. These parameters were used to calculate acoustic $p_{\Lambda}(X|W)$ and linguistic $p_{\Gamma}(W)$ likelihoods to determine the optimal word sequence $\hat{W}$ according to plug-in MAP decoding in (1). Nevertheless, considering the *hierarchical structure* from phonetic-level matching to sentence-level matching, such assumption was unrealistic to estimate truly optimal acoustic and linguistic model parameters. In [13], transition weight and language model in finite state decoding graphs were simultaneously optimized under MCE criterion. In [7], HMM and unigram features were induced for hidden condition random fields (HCRFs) based phone classification. Maximum conditional likelihood criterion was used for parameter estimation. Here, we systematically build an integrated model combining HMM and *n*-gram features using the maximum entropy principle for continuous speech recognition. More importantly, we present the modularized framework for joint estimation of acoustic and linguistic parameters. The dependence between acoustic and linguistic features is properly considered. In the experiments on large vocabulary continuous speech recognition (LVCSR), we illustrate the effectiveness of using integrated ME approach compared to conventional ML training in plug-in MAP classification system.

2. MAXIMUM ENETROPY PRINCIPLE

Statistical modeling approaches are crucial for many pattern recognition applications. Among them, the maximum entropy (ME) principle is very attractive because model parameters are calculated with maximum randomness. In application of speech recognition, ME estimation was first successfully applied for language modeling [6]. Hybrid language models combining *n*-gram, syntactic and semantic regularities were developed [4][10]. Using these models, different linguistic features were integrated delicately. We also exploited a discriminative ME language model incorporating new features from acoustic events [3]. Recently, ME approach was extended to direct modeling [12] of acoustic HMM's. All of these methods achieved desirable model perplexity and recognition accuracy.

Basic idea of ME principle [8] intends to completely model what we observe, and assume nothing about what we do not observe. Using ME approach, all information sources are formulated as constraint sets. Under these constraints, we maximize the entropy and find the optimal model $p(y)$. Namely, we restrict the estimated model to be consistent with all the information sources we have and simultaneously make the model distribution as uniform as possible. Let $f_1, \cdots, f_F$ denote the feature set specifying the properties that we want to integrate in the model. Feature functions can be defined using zero-one delta function

$$f_i(y)=\begin{cases}1 & \text{if } y \text{ matches feature } i \\ 0 & \text{otherwise}\end{cases}. \quad (2)$$

Then, we calculate the expectation of feature functions with respect to empirical distribution $\tilde{p}(y)$ and actual distribution $p(y)$ as follows

$$\tilde{p}(f_i)=\sum_y \tilde{p}(y)f_i(y)=\frac{1}{R}\sum_{i=1}^{R} f_i(y), \quad (3)$$

$$p(f_i)=\sum_y p(y)f_i(y), \quad (4)$$

where R is the number of training samples. Because the actual model encapsulates all information sources, the expectation functions should satisfy the equality

$$\tilde{p}(f_i)=p(f_i), \quad \text{for } i=1,\cdots,F. \quad (5)$$

To solve constrained optimization problem, the Lagrange optimization procedure is applied by merging multipliers $\lambda=\{\lambda_i\}$ in the entropy based objective function

$$H(p,\lambda)=-\sum_y p(y)\log p(y)+\sum_{i=1}^{F}\lambda_i\left[p(f_i)-\tilde{p}(f_i)\right], \quad (6)$$

Through maximizing $H(p,\lambda)$, we yield the ME model expressed in a form of log linear or Gibbs distribution.

$$p(y)=\exp\left(\sum_{i=1}^{F}\lambda_i f_i(y)\right)\Big/\sum_{y'}\exp\left(\sum_{i=1}^{F}\lambda_i f_i(y')\right). \quad (7)$$

To determine Lagrange multipliers, the generalized iterative scaling (GIS) algorithm [5] can be adopted. We briefly describe GIS algorithm below.

Input: Feature functions $f_1,\cdots,f_F$ & empirical distribution $\tilde{p}(y)$

Output: Optimal Lagrange multipliers $\hat{\lambda}$

1. Initialization with $\{\lambda_i=0, i=1,\cdots,F\}$
2. For each $i=1,\cdots,F$, update λ_i based on

$$\lambda_i \leftarrow \lambda_i+\frac{1}{F_i}\log\frac{\tilde{p}(f_i)}{p(f_i)}, \quad (8)$$

$$F_i=\frac{1}{\sum_{y'} p(y')f_i(y')}\sum_y p(y)f_i(y)\sum_{i'} f_{i'}(y). \quad (9)$$

3. Go to step 2 if λ_i has not converged.

After finding optimal parameters $\hat{\lambda}$, we calculate ME model using (7). This ME approach has achieved great success in language modeling [6] $p(y)\to p_\Gamma(W)$ and recently been explored in acoustic modeling [12] $p(y)\to p_\Lambda(X|W)$.

3. ACOUSTIC AND LINGUISTIC MODELING

No matter using ML, MMI, MCE or ME criterion, acoustic and language models were separately estimated for plug-in MAP speech recognition. The dependencies between acoustic and language models were neglected in model estimation. Strictly speaking, HMM and *n*-gram parameters express the information sources in different levels conveying important cues for finding reliable word candidates underlying input speech signal. These two model sets should be jointly established in a consistent way or following the same objective function. To release the independence assumption, one statistically attractive approach is to adopt ME principle simultaneously for acoustic and linguistic modeling. In what follows, we construct a modularized ME framework integrating acoustic and linguistic features for building speech recognition system.

3.1 Joint Acoustic and Linguistic Models

When jointly estimating HMM Λ and *n*-gram Γ parameters, we should consider state S and mixture component L sequences for directly representing the posterior distribution $p_\Theta(W|X)=\sum_{S,L} p_\Theta(W,S,L|X)$ used for MAP classification in (1). New parameters Θ should contain all acoustic and linguistic features characterizing the parametric distributions of HMM's and *n*-grams. According to ME principle, we estimate posterior distribution $p_\Theta(W,S,L|X)$ via maximizing the conditional entropy. ME Lagrangian function is yielded by

$$\begin{aligned}H_{\mathrm{LA}}(p,\lambda)=&-\sum_{W,X,S,L}\tilde{p}(X)p_\Theta(W,S,L|X)\log p_\Theta(W,S,L|X)\\&+\sum_{i=1}^{F}\lambda_i^{\mathrm{LA}}\left[p(f_i)-\tilde{p}(f_i)\right]\end{aligned}. \quad (10)$$

Parameters turn out to be the Lagrange multipliers for *n*-gram λ^{L} and HMM λ^{A}, $\Theta=\lambda^{\mathrm{LA}}=\{\lambda^{\mathrm{L}},\lambda^{\mathrm{A}}\}$. More specifically, we can define the feature functions $f_i^{\mathrm{LA}}(W,X,S,L)$ corresponding to *n*-gram, HMM initial state probability, state transition probability, mixture weight, first-order statistics and second-order statistics of observations as follows

$$f^{\mathrm{L}}_{w_{i-n+1}^{i}}(W,X,S,L)=\begin{cases}\mathrm{count}(w_{i-n+1}^{i}) & \text{if } w_{i-n+1}^{i}\in W\\ 0 & \text{otherwise}\end{cases}, \quad (11)$$

$$f^{\pi}_{s_1}(W,X,S,L)=\delta(S(1)=s_1), \quad (12)$$

$$f^{a}_{s_{t-1}^{t}}(W,X,S,L)=\sum_t \delta(S(t-1)=s_{t-1})\delta(S(t)=s_t), \quad (13)$$

$$f^{\omega}_{s_t l_t}(W,X,S,L)=\sum_t \delta(S(t)=s_t)\delta(L(t)=l_t), \quad (14)$$

$$f^{m_1}_{s_t l_t d}(W,X,S,L)=\sum_t x_{t,d}\,\delta(S(t)=s_t)\delta(L(t)=l_t), \quad (15)$$

$$f^{m_2}_{s_t l_t d}(W,X,S,L)=\sum_t x_{t,d}^2\,\delta(S(t)=s_t)\delta(L(t)=l_t), \quad (16)$$

where $\delta(\cdot)$ is a delta function, and i, t and d are indices of word, frame and dimension, respectively. Using these linguistic and acoustic features $f^{\mathrm{LA}}=\{f^{\mathrm{L}},f^{\pi},f^{a},f^{\omega},f^{m_1},f^{m_2}\}$, ME constraints are setup by

$$\begin{aligned}\tilde{p}(f_i^{\mathrm{LA}})&=\sum_{W,X}\tilde{p}(W,X)\sum_{S,L}p(S,L|W,X)f_i^{\mathrm{LA}}(W,X,S,L)\\&=\sum_{W,X,S,L}p_\Theta(W,X,S,L)f_i^{\mathrm{LA}}(W,X,S,L)=p(f_i^{\mathrm{LA}})\end{aligned}. \quad (17)$$

where $p(S,L|W,X)$ is used for encoding the latent variable based on current estimation because we can not determine the empirical probability $\tilde{p}(W,X,S,L)$ from training data. In this way, EM-IS algorithm [16] is adopted to estimate ME model with hidden variables. To reduce the computation complexity, we assume $\tilde{p}(X)=p_\Theta(X)$ for calculating the expectation function

$$p(f_i^{\mathrm{LA}}) = \sum_{W,X,S,L} \widetilde{p}(X) p_\Theta(W,S,L|X) f_i^{\mathrm{LA}}(W,X,S,L) . \quad (18)$$

Then, we apply these features and constraints for estimation of optimal linguistic λ^{L} and acoustic $\lambda^{\mathrm{A}} = \{\lambda^\pi, \lambda^a, \lambda^\omega, \lambda^{m_1}, \lambda^{m_2}\}$ parameters. Notably, the joint distributions of different parameters are considered in parameter estimation. Dependencies are characterized implicitly. Having all estimated parameters, the integrated ME posterior probability model is generated by

$$p_\Theta(W,S,L|X) = \frac{1}{p_\Theta(X)} \exp \sum_{i=1}^{F} \lambda_i^{\mathrm{LA}} f_i^{\mathrm{LA}}(W,X,S,L) , \quad (19)$$

with normalization term $p_\Theta(X)$. Here, we directly estimate the posterior probability using ME approach so that HMM and *n*-gram parameters can be optimized simultaneously.

3.2 Relations of Feature Functions to HMM and N-Gram

Feature functions are used to determine the scope of the knowledge sources. It is desired to extract knowledge from HMM's and *n*-gram for ME modeling. We would like to expand posterior distribution to express the specification of feature functions for HMM and *n*-gram features. Using continuous-density HMM parameters $\{\pi, A, B\} = \{p(s_1), p(s_t|s_{t-1}), p(l_t|s_t), \mu_{s_t l_t d}, \sigma^2_{s_t l_t d}\}$ and *n*-gram parameters $\{p(w_i|w_{i-n+1}^{i-1})\}$, the posterior probability is expressed as

$$\begin{aligned} p(W,S,L|X) &= (p(W)p(S,L|W)p(X|S,L,W)) / p(X) \\ &= \Bigg(\prod_i p(w_i|w_{i-n+1}^{i-1}) \cdot p(s_1) \cdot \prod_t p(s_t|s_{t-1}) p(l_t|s_t) \cdot \\ &\quad \prod_d \frac{1}{\sqrt{2\pi\sigma^2_{s_t l_t d}}} \exp\left(-\frac{(x_{t,d} - \mu_{s_t l_t d})^2}{2\sigma^2_{s_t l_t d}} \right) \Bigg) \Big/ p(X) \end{aligned} \quad (20)$$

After careful arrangement and neglecting normalization term $p(X)$, this posterior distribution can be represented in a consistent form consisting of different sets of log linear distributions

$$\begin{aligned} \exp\Bigg(& \sum_{w_{i-n+1}^i} \lambda^{\mathrm{L}}_{w_{i-n+1}^i} f^{\mathrm{L}}_{w_{i-n+1}^i}(W,X,S,L) + \sum_{s_1} \lambda^\pi_{s_1} f^\pi_{s_1}(W,X,S,L) \\ & + \sum_{s_{t-1},s_t} \lambda^a_{s_{t-1}^t} f^a_{s_{t-1}^t}(W,X,S,L) + \sum_{s_t,l_t} \lambda^\omega_{s_t l_t} f^\omega_{s_t l_t}(W,X,S,L) \\ & + \sum_{s_t,l_t,d} \lambda^{m_1}_{s_t,l_t,d} f^{m_1}_{s_t,l_t,d}(W,X,S,L) + \sum_{s_t,l_t,d} \lambda^{m_2}_{s_t,l_t,d} f^{m_2}_{s_t,l_t,d}(W,X,S,L) \Bigg) \end{aligned} . \quad (21)$$

In this way, we derive the feature functions in (11)-(16). The information sources of HMM and *n*-gram can be properly extracted and incorporated in the integrated ME model.

3.3 Implementation Issues

In our implementation, we adopt N-best lists to approximate the denominator term in the objective function when computing the posterior probability. Viterbi alignment is performed to find the optimal state $\hat{S}$ and mixture component $\hat{L}$ sequences corresponding to each word sequence in N-best list. Interestingly, ME objective function can be also represented as

$$H_{\mathrm{LA}} \approx \sum_{W,X} \widetilde{p}(W,X) \log \frac{\exp \sum_i \lambda_i^{\mathrm{LA}} f_i^{\mathrm{LA}}(W,X,\hat{S},\hat{L})}{\sum_{W_c \in M_W} \exp \sum_i \lambda_i^{\mathrm{LA}} f_i^{\mathrm{LA}}(W_c,X,\hat{S},\hat{L})} . \quad (22)$$

where M_W is the competing set of word sequence *W*. By partially differentiating (22) with respect to λ_i^{LA}, the actual expectation function can be approximated by considering all possible word sequences in N-best list

$$\begin{aligned} p(f_i^{\mathrm{LA}}) &\approx \sum_{W,X} \widetilde{p}(W,X) \sum_{W_c \in M_W} \frac{p_\Theta(W_c,\hat{S},\hat{L}|X)}{\sum_{W'_c \in M_W} p_\Theta(W'_c,\hat{S},\hat{L}|X)} f_i^{\mathrm{LA}}(W_c,X,\hat{S},\hat{L}) \\ &= p_{\mathrm{N-best}}(f_i^{\mathrm{LA}}) \end{aligned} \quad (23)$$

Using this new ME constraints, the learning rule of Lagrange multipliers in EM-IS algorithm can be updated by

$$\lambda_i^{\mathrm{LA}} \leftarrow \lambda_i^{\mathrm{LA}} + \frac{1}{F_i} \log\left(\frac{\widetilde{p}(f_i^{\mathrm{LA}})}{p_{\mathrm{N-best}}(f_i^{\mathrm{LA}})} \right) . \quad (24)$$

The implementation procedure of proposed method is shown in Figure 1. The training speech data are first recognized to extract competing information using a seed HMM and *n*-gram models. Through EM-IS algorithm, the acoustic and linguistic features are integrated. We finally estimate the integrated acoustic and linguistic ME models for speech recognition. Generally, the proposed integrated ME is similar to HCRFs [7] regarding the joint acoustic and language modeling. However, in HCRFs, model distribution is defined using a Gibbs distribution with parameters λ estimated by maximizing the conditional log-likelihood of training data. The integrated ME modeling performs constrained optimization without specifying the form of model distribution. The model with maximum uncertainty is selected.

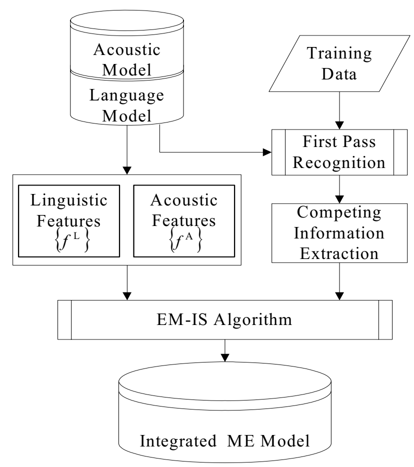

Figure 1. Implementation procedure of integrated ME model.

4. EXPERIMENTAL RESULTS

In the experiments, we estimated the seed speaker-independent HMM models using the benchmark Mandarin speech corpus TCC300, which was recorded in office environment using close-talking microphone. Each Mandarin syllable was modeled by right context-dependent states with at almost 32 mixture components. We trained sub-syllable HMMs for Mandarin speech. Feature vectors consisted of twelve Mel-frequency cepstral coefficients, one log energy and their first derivation. Here, another speech corpus, Mandarin Across Taiwan Broadcast News (MATBN) was applied for environmental adaptation adjusting acoustic models from TCC300 to MATBN database via maximum *a posteriori*

(MAP) adaptation. MATBN database contained a total of 198 hours of broadcast news from the Public Television Service Foundation (Taiwan) and was arranged by Academic Sinica, Taiwan. All recordings were made in stereo with a 44.1 kHz sampling rate and 16 bit resolution via a single channel microphone. Signals were down-sampled to 16 kHz. Totally, we selected 1060 conversations about 270 minutes for acoustic model adaptation, and another 250 sentences about 30 minutes for testing. The adaptation data was also used to extract N-best list. In our experiments, only top one competing sentence was explored. Using Katz back-off smoothing, baseline ME bigram language model was trained using Academic Sinica CKIP balanced corpus constructed by about five million words with lexical size of 32,909.

In the experiments, we focused on three major parameters, including *n*-gram, state transition (Trans.) and first-order observation parameters (Obs_1.). To evaluate speech recognition performance, we report syllable, character and word error rates. In Table 1, we show the performance with individually updated parameters. Baseline system adopts conventional HMM and *n*-gram parameters.

Table 1. Evaluation for different updated parameters

	Baseline	Maximum Entropy		
		N-gram	Trans.	Obs_1.
SER (%)	29.1	28.6 (1.7)	28.6 (1.7)	28.5 (2.2)
CER (%)	36.8	35.3 (4.1)	36.4 (1.1)	35.9 (2.4)
WER (%)	48.9	46.9 (4.1)	48.3 (1.2)	47.8 (2.2)

Interestingly, we obtain desirable syllable error rates via updating observation parameters while good word error rates are achieved by updating *n*-gram parameters. This is reasonable because observation parameters play the critical role for modeling acoustic properties, which affects syllable error rate significantly. *N*-gram parameters are used to characterize word association so that word accuracy can be improved.

Table 2. Evaluation for combinations of updated parameters

	Baseline	Maximum Entropy			
		Trans.+ Obs_1.	*N*-gram+Trans.	*N*-gram+Obs_1.	All
SER (%)	29.1	28.1(3.4)	28.0(3.9)	28.0(3.9)	27.8(4.5)
CER (%)	36.8	35.7(3.0)	35.1(4.6)	34.9(5.2)	34.6(6.0)
WER (%)	48.9	47.6(2.7)	46.5(4.9)	46.5(4.9)	46.0(5.9)

We also implemented the experiments using integrated model to illustrate the effectiveness of different parameter combinations. Error rates using different parameter combinations are shown in Table 2. We find that integrated parameters achieve better improvements than using individually updated parameters. Updated *n*-gram parameters have slight improvement of syllable error rate with 1.7% and combined *n*-gram and acoustic features can achieve 3.9%. Namely, combining language and acoustic models in ME model does improve speech recognition. Furthermore, we obtain the best word error rate improvement with 5.9% based on integrated model with all updated parameters.

5. CONCLUSIONS

We have presented a joint modeling approach to acoustic and linguistic features to effectively estimate correlated parameters for speech recognition. This approach released the independence assumption which was made by many speech recognition systems. Importantly, we developed a ME framework jointly merging HMM and *n*-gram features and modeling their dependencies in a consistent and optimal fashion. Specially, ME model involved a constrained optimization procedure and was powerful for knowledge integration. In the experiments on broadcast news transcription, we obtained desirable performance compared to conventional recognition system using independent HMM and *n*-gram parameters.

6. REFERENCES

[1] L. Bahl, P. Brown, P. de Souza and R. Mercer, "Maximum mutual information estimation of hidden Markov model parameters for speech recognition", in *Proc. of ICASSP*, vol. 1, pp. 49-52, 1986.

[2] Z. Chen, K.-F. Lee, M.-J. Li, "Discriminative training on language model," in *Proc. of ICSLP,* pp. 16-20, 2000.

[3] C.-H. Chueh, T.-C. Chien, and J.-T. Chien, "Discriminative maximum entropy language model for speech recognition," in *Proc. of INTERSPEECH*, pp. 721-724, 2005.

[4] C.-H. Chueh, J.-T. Chien and H. Wang, "A maximum entropy approach for integrating semantic information in statistical language models", in *Proc. of ISCSLP*, pp. 309-312, 2004.

[5] J. Darroch and D. Ratcliff, "Generalized iterative scaling for log-linear models", *The Annals of Mathematical Statistics*, vol. 43, pp. 1470-1480, 1972.

[6] S. Della Pietra, V. Della Pietra, R. L. Mercer and S. Roukos, "Adaptive language modeling using minimum discriminant estimation", in *Proc. of ICASSP*, vol. 1, pp. 633-636, 1992.

[7] A. Gunawardana, M. Mahajan, A. Acero and J. C. Platt, "Hidden conditional random fields for phone classification", in *Proc. of INTERSPEECH*, pp. 1117-1120, 2005.

[8] E. Jaynes, "Information theory and statistical mechanics", *Physics Reviews*, vol. 106, no. 4, pp. 620-630, 1957.

[9] B.-H. Juang and S. Katagirl, "Discriminative learning for minimum error classification", *IEEE Trans. Signal Processing*, vol. 40, pp. 3043-3054, 1992.

[10] S. Khudanpur and J. Wu, "Maximum entropy techniques for exploiting syntactic, semantic and collocational dependencies in language modeling", *Computer Speech and Language*, vol. 14, pp. 355-372, 2000.

[11] H.-K. J. Kuo, E. Fosle-Lussier, H. Jiang and C.-H. Lee, "Discriminative training of language models for speech recognition", in *Proc. of ICASSP*, vol. 1, pp. 325-328, 2002.

[12] H.-K. J. Kuo and Y. Gao, "Maximum entropy direct model as a unified model for acoustic modeling in speech recognition", in *Proc. of ICSLP*, 2004.

[13] S.-S. Lin and F. Yvon, "Discriminative training of finite state decoding graphs", in *Proc. of INTERSPEECH*, pp. 733-736, 2005.

[14] Y. Normandin, R. Cardin and R. De Mori, "High-performance connected digit recognition using maximum mutual information estimation", *IEEE Trans. Speech and Audio Processing*, vol. 2, pp. 299-311, 1994.

[15] R. Rosenfeld, "A maximum entropy approach to adaptive statistical language modeling", *Computer Speech and Language*, vol. 10, pp. 187-228, 1996.

[16] S. Wang, D. Schuurmans, F. Peng and Y. Zhao, "Learning mixture models with the regularized latent maximum entropy principle", *IEEE Trans. Neural Networks*, vol. 15, no. 4, 2004.