A TOOL FOR AUTOMATED DESIGN OF LANGUAGE MODELS

Y.P. Yang & J.R. Deller, Jr.

Michigan State University
Department of Electrical Engineering / Speech Processing Laboratory
East Lansing, MI 48824-1226 USA
{yangy, deller}@egr.msu.edu

ABSTRACT

An interactive software tool for design and performance analysis of language models (LMs) is described. The tool obviates on-line simulation of the recognition system in which the LM is to employed. By exploiting parallels with signal detection theory, a profile of the LM is given in an receiver-operating-characteristic-like (ROC) display.

1. INTRODUCTION

"ROC-LM" Design Tool. The benefits of LMs in modern speech recognition (SR) systems are well-understood (e.g. [1, Ch. 13]). Many classes of LMs, including adaptive (e.g. [2, 4]) and hybrid (e.g. [3]) languages, have been researched, but few simple models, notably n-grams, prevail. A serious impediment to the development and testing of LMs is the inability to assess their performance without time-consuming experimentation. In this paper, we develop a new LM design / evaluation tool ("ROC-LM Design Tool") that does not require simulation with the actual SR system.

The ROC-LM tool exploits parallels between binary signal detection theory and the LM problem. An LM must be "large" to "cover" the task language (TL). A "large" LM assures proper "detection" of a TL sentence. Conversely, a large LM tends to "over-generate" out-of-TL hypotheses, risking "false alarms" (acceptance of out-of-TL sentences). For a fixed performance criterion (e.g., recognition rate) analogous to signal-to-noise ratio in detection theory, and for a fixed set of acoustic models, contours of "coverage / over-generation" pairs, (C_M, O_M), can be mapped in a two-dimensional ROC-like plane [5]. A point on a contour applies to *any* LM that is characterized by the given pair. For a specific LM of interest, the designer can monitor performance in the plane as certain LM design features are varied.

Limitations of the Current System. The present version of the ROC-LM tool has practical limitations that will be alleviated by ongoing research. The most limiting present assumption is that word (or other linguistic) boundaries are known in the acoustic observations. Other *theoretical* assumptions (for which the practical implications are not yet entirely understood) are the assumption of constant sentence length in the TL, and the assumption of uniform distributions of elements in the TL and LM. These later assumptions can certainly be relaxed through further theoretical work (see [6]). Other assumptions (validated by current research) regarding statistics of acoustic likelihoods are described below.

2. FORMULATION

General SR Problem. The general SR task can be viewed as formal signal detection problem as illustrated in Fig. 1 [5]. The *vocabulary*, V, is an exhaustive set of N_V words, including a "null" word, that arises in the TL. If the maximum sentence length is $l_{\max}$, there are $\mathcal{O}\left(N_L = N_V^{l_{\max}}\right)$ possible sentences (word strings), say, $L = \{s_1, \ldots, s_{N_L}\}$. For each s_i, there is a (assumed finite) set of acoustic representations ("pronunciations"), denoted $A = \{a_1, \ldots, a_{N_A}\}$. The pronunciations and acoustic strings arising from frame procesing are in one-to-one correspondence and need not be distinguished.

The underlying probability space is $(\Omega, 2^\Omega, P)$, where P reflects the "true" (TL) distribution of pairs $\omega_{ij} = (s_i, a_j) \in \Omega$. Formally, when a sentence s_i is "pronounced as" a_j, $\omega_{ij} \in \Omega$ is selected; this occurs with probability $P(\omega_{ij})$. Outcome ω_{ij} is then analyzed by the "transmitter" using a *source classifier*[1] $\underline{c}_R(\omega_{ij}) = s_i$. At the "receiver" (recognizer), the *observation* $\underline{o}(\omega_{ij}) = a_j$ is made. Based on $\underline{o}$, *decision rule* is

$$\underline{\delta}_R(a_j) = \arg\max_k \{-\ln Q(a_j|s_k)Q(s_k)\} = s_{\hat{\imath}}, \qquad (1)$$

where $Q(\cdot)$ is defined below. Let l denote the decision statistic used in $\underline{\delta}_R$,

$$\begin{aligned} l(s_k, a_j) &= -|s_k|^{-1}|a_j|^{-1}\ln Q(a_j|s_k)Q(s_k), \quad \text{or,} \\ l(s_i, a_j) &= l_a(a_j|s_k) + l_g(s_k) \qquad (2) \\ &= \left[-|a_j|^{-1}\ln Q(a_j|s_k)\right] + \left[-|s_i|^{-1}\ln Q(s_k)\right], \end{aligned}$$

where $|a_j|$ [$|s_i|$] denotes the length of the acoustic string [sentence] in frames [words]. In (2), l is decomposed into negative log likelihoods associated with the acoustic and language

[1] The underscore indicates a random variable. The purist should read "$\underline{c}_R = s_i$," e.g., as "$\underline{c}_R = i$" so that $\underline{c}_R$ has an appropriate domain.

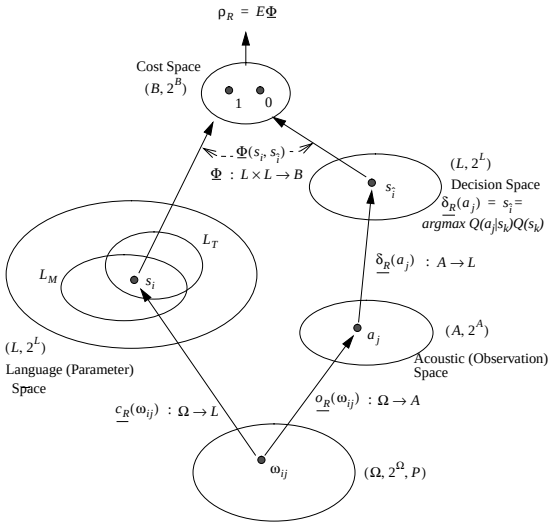

Figure 1: SR viewed as a formal signal detection problem.

models (subscript on l_g denotes "grammar"). Finally, the *cost function* $\underline{\Phi}$ determines the correctness of the outcome: $\underline{\Phi}(s_i, s_{\hat{i}}) = \{1, i = \hat{i};\ \ 0, i \neq \hat{i}\}$. The *recognition rate* is given by $\rho_R = E\underline{\Phi}$. This rate, or further cost criteria, may be used to adjust the LM as illustrated below.

The mapping Q may be viewed as a "noisy probability distribution" over Ω learned in the training process. It represents an attempt to estimate the TL "true" distribution P. In the language space L, there are two critical subsets associated with these distributions. Formally, the *task language* (TL), $L_T \subseteq L$, is defined as the subset $L_T = \{s_i | P(s_i) \neq 0; s_i \in L\}$, whereas the *language model* (LM), $L_M \subseteq L$, is $L_M = \{s_i | Q(s_i) \neq 0; s_i \in L\}$. Therefore $P(L_T) = 1$ and $Q(L_M) = 1$. For convenience, we also define subsets $L_{MT} = L_M \cap L_T$, $L_{\bar{T}} = L_T - L_{MT}$, and $L_{\bar{M}} = L_M - L_{MT}$ (Fig. 2).

To avoid tedious arguments, we henceforth assume that P and Q are approximately uniform over their support sets. Generalizations are discussed in [6] and are the subject of ongoing research.

Language Detection (LD) Problem. The heart of the speech recognition system is the detection rule δ_R. In turn, the "performance" of the rule δ_R reflects the quality of the training as reflected in the distribution Q. The recognition performance can be characterized in terms of the question "How well does Q approximate P over the space $\Omega = L \times A$?". In this work, however, we are interested in the quality of the LM, and this part of the training is reflected in the distribution $Q(s_i)$ over the subspace L and the corresponding term $Q(s_i)$, or $l_g(s_i)$, in δ_R. The "LD problem" which is structured similarly to Fig. 1, is a *formal construct only*, allowing the quantified study of the LM as reflected in $Q(s_i)$ and δ_R.

In the LD problem, the fundamental probability space is $(L, 2^L, \hat{P})$ where L is as defined above, and $\hat{P}$ is uniform over *all* $s_i \in L$. The distributions P and Q induced on L in the recognition problem (Fig. 1) are such that $P(s_i) = \hat{P}(s_i|L_T)$ and $Q(s_i) = \hat{P}(s_i|L_M)$ for any $s_i \in L$, where the subsets L_T and L_M are also as defined above. In the LD problem, a test sentence s_i can be selected from anywhere in L, with probability $\hat{P}(s_i)$, to evaluate the LM. In detection theory terms, the test sentence s_i is then analyzed by the "transmitter" using a *source classifier*: $\underline{c}_L(s_i) = \underline{I}_T(s_i)$, where $\underline{I}_T$ is the indicator mapping for set L_T. At the "receiver" (detector), the *observation* is the sample sentence itself: $\underline{o}_L(s_i) = s_i$. The *decision rule* is $\underline{\delta}_L(s_i) = \underline{I}_M(s_i)$ where $\underline{I}_M$ is the indicator for L_M.[2]

The LD problem is equivalent to a classical "binary channel" problem (e.g. [5, p. 9]) with the following hypothesis structure (letting $s_i \in L$ be a test sentence in a given trial):

$$H_0 : \underline{c}_L = 0 \quad (s_i \notin L_T) \quad \text{and} \quad H_1 : \underline{c}_L = 1 \quad (s_i \in L_T), \tag{3}$$

and the following decision structure:

$$D_0 : \underline{\delta}_L = 0 \quad (s_i \notin L_M) \quad \text{and} \quad D_1 : \underline{\delta}_L = 1 \quad (s_i \in L_M). \tag{4}$$

The "channel transition" probabilities associated with this problem are: $\hat{P}(D_1|H_1) = \hat{P}(\underline{\delta}_L = 1|\underline{c}_L = 1)$ (conditional LD rate (hit rate), related to "coverage" below); $\hat{P}(D_1|H_0) = \hat{P}(\underline{\delta}_L = 1|\underline{c}_L = 0)$ (conditional false-alarm rate, related to "over-generation" described below); $\hat{P}(D_0|H_1) = \hat{P}(\underline{\delta}_L = 0|\underline{c}_L = 1)$ (conditional miss rate); and $\hat{P}(D_0|H_0) = \hat{P}(\underline{\delta}_L = 0|\underline{c}_L = 0)$ (conditional correct-rejection rate). Figure 2 illustrates the set relationships among these four probabilities in the space L. These channel probabilities capture the "performance" of the distribution Q with respect to the LM, since it is upon Q (via $\underline{\delta}_R$) that the decisions are based.

The analysis of the LM model is fundamentally based on the ROC-like characterization of the LD problem. The impetus for using ROC analysis in LM design derives from an interesting trade-off phenomenon. A LM needs to be "flexible" (L_M sufficiently large) so that the complete complement of sentences can be hypothesized during recognition. However, if the model is so flexible ($N_M >> N_T$) that many sentences outside of L_T are hypothesized, the computational cost is increased and the recognition rate is decreased.

In these heuristic terms, the performance of a LM is parameterized by "coverage" and "over-generation." The *coverage*,

[2] For completeness (although the "output step" is not used in the following developments): The *cost function* is the product $\underline{\phi}_L(c_L, \delta_L) = c_L \times \delta_L$, and the expectation of $\underline{\phi}_L$ with respect to the distribution $\hat{P}$ is the *LD rate*, $\hat{\rho}_L = \hat{E}\underline{\phi}_L = \hat{P}(L_{MT})$.

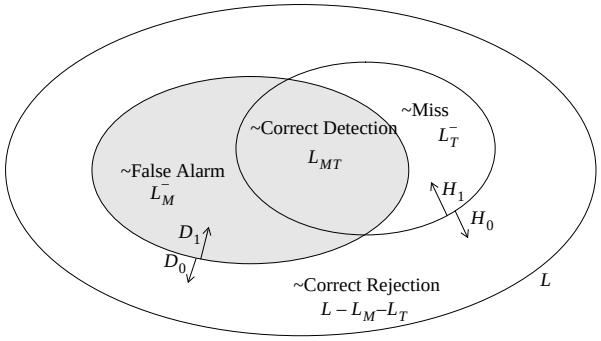

Figure 2: Illustration of the four channel probabilities, conditional LD, false-alarm, miss, and correct-rejection, rates.

C_M, of the LM is exactly equivalent to the conditional hit rate of the LD problem and it reflects the probability that a sentence in the TL will be hypothesized ("covered" in the sense of Fig. 2) by the LM. Note, therefore, that

$$C_M \stackrel{\text{def}}{=} \hat{P}(D_1|H_1) = \hat{P}(L_M|L_T) = P(L_M) = P(L_{MT}). \quad (5)$$

The *over-generation*, O_M, is defined to be the "*a posteriori* false-alarm rate"

$$O_M \stackrel{\text{def}}{=} \hat{P}(\underline{c}_L = 0|\underline{\delta}_L = 1) = \hat{P}(H_0|D_1). \quad (6)$$

O_M differs from the proper false-alarm rate $\hat{P}(D_1|H_0)$ only by a scale factor, which, in turn, avoids the occurrence of very small numbers in the analysis. O_M, so-defined, is a normalized measure of the size of the LM.

Integrating the LD and SR Problems. Having achieved the goal of defining the key measures C_M and O_M, we abandon the use of distribution $\hat{P}$ and return to the exclusive use of P and its estimate Q. Using the developments above, it can be shown that [6]

$$\rho_R = C_M \sum_{s_i \in L_{MT}} P(\underline{\delta}_R = s_i|\underline{c}_R = s_i) P(\underline{c}_R = s_i|L_{MT}) \quad (7)$$

Now let Λ_i be the set of acoustic strings corresponding to sentence s_i, $\Lambda_i \stackrel{\text{def}}{=} \{a_j|P(a_j, s_i) > 0\}$, and let $\bar{\Lambda}_i$ be the set of acoustic strings *not* corresponding to sentence s_i, $\bar{\Lambda}_i \stackrel{\text{def}}{=} \{a_j|P(a_j, s_i) = 0\}$. By experimentation, we have determined that the decision statistics[3]

$$\underline{l}_i(a_j) = -|a_j|^{-1}|s_i|^{-1} \ln Q(a_j|s_i, a_j \in \Lambda_i) Q(s_i), \quad (8)$$

are approximately independent and identically distributed (i.i.d.) over $s_i \in L_{MT}$. We denote the distribution of each of these random variables by $f(\cdot)$. Likewise, for $s_i \in L_{MT}$ and $s_i \in L_M^-$, respectively, the statistics

$$\begin{aligned} \bar{\underline{l}}_{i,MT}(a_j) &= -|a_j|^{-1}|s_i|^{-1} \ln Q(a_j|s_i, a_j \in \bar{\Lambda}_i) Q(s_i) \\ \bar{\underline{l}}_{i,-}(a_j) &= -|a_j|^{-1}|s_i|^{-1} \ln Q(a_j|s_i, a_j \in \bar{\Lambda}_i) Q(s_i) \end{aligned}$$

[3] These results do not depend on the form of Q; however, the assumed uniformity of Q in the present work implies that the factor $Q(s_i)$ in each of the statistics below becomes superfluous and may be deleted.

are i.i.d. with distributions $\bar{f}_{L_{MT}}(\cdot)$ and $\bar{f}_{L^-}(\cdot)$. The estimation of the distributions f, $\bar{f}_{L_{MT}}$, and $\bar{f}_{L^-}$ using combinations of Gaussian densities from real data is a tedious procedure described in [6]. It is assumed (without a complete understanding of practical consequences) that $|s_i|$ (sentence length) is constant over L_T. Details and remarks on relaxing this assumption are found in [6].

In these terms, the recognition rate can be expressed as [6]

$$\rho_R = C_M \int_0^\infty f(x) \left[\int_x^\infty \bar{f}_{L^-}(\bar{x}) d\bar{x}\right]^\alpha \left[\int_x^\infty \bar{f}_{L_{MT}}(\bar{x}) d\bar{x}\right]^\beta dx, \quad (9)$$

with $\alpha \stackrel{\text{def}}{=} N_T(C_M O_M)/(1 - O_M)$ and $\beta \stackrel{\text{def}}{=} (C_M N_T - 1)$. ρ_R is the performance measure used in assessing the quality of a LM below. Note that this figure is conservative in expressing the percentage of *perfectly recognized* sentences. More liberal measures, such as rate of sentences with substitutions or deletions, are similarly derived.

3. THE ROC-LM DESIGN TOOL

The ROC-LM Design Tool is segmented into two interactive components. The first is concerned with the recognition environment including the performance of the acoustic models and the properties of the TL. These factors are assumed fixed once the LM assessment begins. The second component is concerned with the LM itself.

In the ROC-LM operation, a point in the "ROC-LM plane" represents a (C_M, O_M) pair, hence a nominal LM with respect to a nominal TL. (An example screen output, discussed below, appears in Fig. 4.) Performance measures [in the present case, ρ_R using (9)] are estimated for a set of grid points on the ROC-LM plane. Contour plots parameterized by these performance measures are then drawn on the chart. Any specific LM to be employed with the given acoustic models can then be evaluated by estimating and plotting its (C_M, O_M) pair in the plane. The designer can easily observe the "motion" of the LM in the plane when its design factors are modified.

An X-Window version of the ROC-LM Design Tool has been implemented on a Sun Workstation SPARC 10 with MATLAB graphics support. Figure 3 shows a diagram of the design tool. Figure 4 illustrates a typical output.

4. EXAMPLE APPLICATION

As an example, we consider an adaptive, hybrid, LM design problem. In this problem a "partial" (or "local") document consisting of the "small" training set $H_{\text{loc}} \in L_T$ does not contain enough sentences to robustly train the LM. Therefore, the larger "static" corpus $H_{\text{st}} \in L_T$ is used to adaptively supply more information. The number of training sentences from H_{st} represents a parameter of the LM design that must be optimized with respect to the LM performance goal. Increasing the amount of H_{st} used will increase the C_M of a

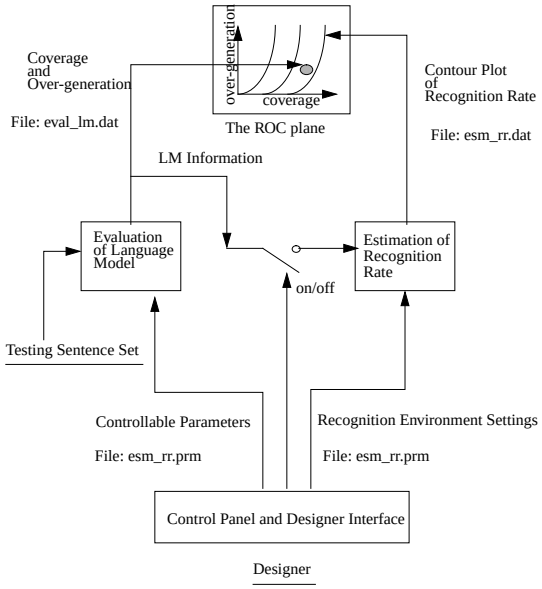

Figure 3: Four blocks of the ROC-LM Design Tool: Display window, Recognition rate evaluator, LM evaluator, and the Control panel & designer interface.

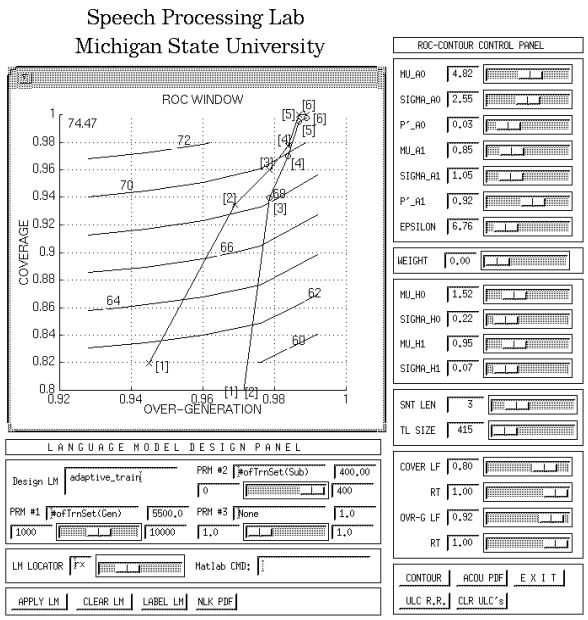

Figure 4: ROC-LM Design Tool output for the application in Section 4. Details are found in the text.

given LM while also increasing its O_M. The TL in this example is an artificial digit-string language described in [1, Ch. 10, Problem 10]. $H_{\rm st}$ is nominally the entire TL, while $H_{\rm loc}$ is just some portion of the "childrens' " partition of the TL. The two LM design parameters ("PRM" on the chart) are the number of $H_{\rm st}$ sentences used (PRM #1) and the number of number of $H_{\rm loc}$ sentences used (PRM #2). Sentences from $H_{\rm loc}$ are used to train both the bi- and tri-gram parts of a hybrid bi-tri-gram LM while those from $H_{\rm st}$ train the bi-gram portion only. The likelihood l_g used in recognition is a weighted average of the bi- and tri-gram outcomes.

The salient features of the ROC-LM output screen for this problem are seen in Fig. 4. Each of the contours shows sets of (C_M, O_M) pairs that result in a constant recognition rate for the acoustic models (whole-word digit HMMs based on the "jojo" database [1, p. ix]) used in the experiments. Each of the two design traces moving toward the "northeast" in the chart represents increasing amounts of $H_{\rm st}$ used in training (PRM #1 increasing in six steps: 1000, 1900, 2800, 3700, 4600, 5500 sentences) for a fixed $H_{\rm loc}$ size (PRM #2 = 0 in the rightmost trace, 400 in the left). With either $H_{\rm loc}$ condition, adding training data from $H_{\rm st}$ improves C_M and decrements O_M, but with a nearly continual improvement in overall recognition performance. (On the second trajectory, the recognition rate reaches its maximum when 3700 $H_{\rm st}$ sentences are used.) The entire display is computed in near-real time without any actual recognition experiments performed.

REFERENCES

1. J.R. Deller, Jr., J.G. Proakis, and J.H.L. Hansen. *Discrete-Time Processing of Speech Signals.* Prentice-Hall / Macmillan, New York, 1993.
2. R. Lau, R. Rosenfeld, and S. Roucos. Trigger-based language models: A maximum entropy approach. In *Proc. IEEE ICASSP '93*, pages II–45 – II–48, 1993.
3. M. Meeter and J.R. Rohlick. Statistical language modeling combining n-gram and context-free grammars. In *Proc. IEEE ICASSP '93*, pages II–37 – II–40, 1993.
4. S.D. Pietra *et al.* Adaptive language modeling using minimum discriminant information. In *Proc. IEEE ICASSP '92*, pages I–633 – I–636, 1992.
5. H.V. Poor. *An Introduction to Signal Detection and Estimation (2d Ed.).* Springer-Verlag, New York, 1994.
6. Y.P. Yang. *Automated Evaluation of Language Models Based on Receiver-Operator-Characteristic Analysis.* PhD thesis, Michigan State U., East Lansing, 1996.

Acknowledgements. The authors are grateful to Professors M. Nayeri, M. Siegel, and J. Hall for valuable inputs. This work was supported in part by an Ameritech Faculty Fellowship.