

K. Kacioglu, W. Ward, "A Figure of Merit for the Analysis of Spoken Dialog Systems", ICSLP-2002:Inter. Conf. on Spoken Language Processing, vol. 2, pp. 877-880, Denver, CO USA, Sept. 2002

# A Figure of Merit for the Analysis of Spoken Dialog Systems



**Kadri Kacioglu, Wayne Ward**

**Center for Spoken Language Research**  
University of Colorado Boulder, Campus Box 594  
3215 Marine Street  
Boulder, Colorado 80309-0594  
<http://cslr.colorado.edu/>



*ICSLP-2002: Inter. Conf. on Spoken  
Language Processing  
INTERSPEECH 2002  
Denver, Colorado USA  
September 16-20, 2002*

# A FIGURE OF MERIT FOR THE ANALYSIS OF SPOKEN DIALOG SYSTEMS

*Kadri Hacioglu and Wayne Ward*

Center for Spoken Language Research  
University of Colorado at Boulder  
E-Mail: {hacioglu,whw}@cslr.colorado.edu

## ABSTRACT

In this paper, a single metric, which we will call the figure of merit, for the quantitative analysis and comparison of spoken dialog systems is introduced. This figure of merit is the product of the weighted dialog accuracy (expressed as the rate of success) and the weighted dialog efficiency (expressed as the average number of concepts per turn). Actually, it is highly desirable to have a quick and accurate dialog. However, these two requirements are conflicting. That is, an improvement in efficiency is accomplished at the expense of accuracy or vice versa. This makes difficult to compare two different spoken dialog systems or tune a particular system. We believe that this figure of merit would avoid those difficulties. To illustrate its use, we consider spoken dialog systems with different dialog strategies and compare them by performing quantitative analysis based on the finite state models of information items using the proposed metric.

## 1. INTRODUCTION

The ultimate goal in spoken dialog systems (SDSs) is to maximize the user satisfaction. However, it is very difficult, if not impossible, to quantify it and use for system optimization. However, it is very common to use the average number of turns for dialog completion along with the probability of success to quantify the performance of a dialog system, as they are believed to be fairly strong predictors of the user satisfaction and can be easily quantified. Within the PARADISE framework [1], the probability of success corresponds to the dialog success which has to be maximized and the average number of turns corresponds to a dialog cost which has to be minimized. We are interested in combining those two metrics into one single metric to make easier the comparison of the systems or the process of optimization, as they are two conflicting metrics.

In this paper, we introduce a metric which we refer to as the figure of merit (FOM). It is the product of the weighted

performance metrics mentioned earlier. The weights are selected to reflect the contribution of each component to the user satisfaction. By doing so, we end up with a metric which is highly correlated with the user satisfaction and very convenient for a quantitative analysis.

The organization of the paper is as follows. In section 2, we introduce the figure of merit. Next, in section 3, we describe several simple spoken dialog systems to be assessed. Quantitative assessment of the systems and comparison of the metrics are numerically presented in section 4. Concluding remarks are made in the last section.

## 2. A FIGURE OF MERIT

Let us denote the probability of success and the average number of system's turns as  $P_s$  and  $N_t$ , respectively. We define the figure of merit as

$$FOM = \left(\frac{n_i}{N_t}\right)^{\lambda_1} P_s^{\lambda_2} \quad (1)$$

where  $n_i$  is the number of information items required by the system to complete a transaction. The weights  $\lambda_1$  and  $\lambda_2$  have to be determined so that they reflect how critical each component is to the ultimate system performance (here, user satisfaction). Note that the first term in  $FOM$  is actually the efficiency of the system.

Based on the assumptions given in Section 3, the ideal case corresponds to  $FOM = 1.0$ ;  $P_s = 1.0$  and  $N_t = n_i$ . That is, in this study, the  $FOM$  in equation (1) is lower-bounded by zero and upper-bounded by one. So, the closer the metric is to 1.0 the better the system is. This metric being bounded to the interval  $[0,1]$  makes the interpretation, optimization, and comparison among different systems easier.

The respective metrics have been correlated to user satisfaction in a linear regression framework (along with additional metrics) in [2]. The weights roughly show that the accuracy of an SDS is on average three times more important than the efficiency of an SDS. In the following, therefore, we set  $\lambda_1 = \frac{1}{3}$  and  $\lambda_2 = 1$ . Indeed, the actual values of the

---

The work is supported by DARPA through SPAWAR under grant #N66001-00-2-8906.

weights should be determined by correlating the FOM in (1) to user satisfaction.

In this paper, the dialog goal,  $G$ , corresponds to

$$G = \{FOM > FOM^*\} \quad (2)$$

where  $FOM^*$  is the minimum acceptable value of the figure of merit. It is the set of system operating points that satisfy the condition given. We are interested in mapping  $G$  on to the plane determined by the confidence estimation metrics and compare it to those obtained in [3] using the component metrics separately.

### 3. DESCRIPTION OF EXAMPLE SYSTEMS

We describe three systems. One is the baseline system which does not use confidence. The other two systems use confidence annotation differently. We first introduce our assumptions:

- SDS needs  $n_i$  information items to complete a transaction,
- Information items are asked one at a time and are independent,
- The user is very cooperative and replies to what has been asked
- The user replies "yes" or "no" to verification
- Speech understanding and confidence annotation performance metrics are the system's overall average

As introduced in [4] an information item can be in one of the following states:

1. unknown;  $(u, \phi)$
2. known, not confirmed, incorrect;  $(k, e)$
3. known, not confirmed, correct;  $(k, c)$
4. accepted, correct;  $(a, c)$
5. accepted, incorrect;  $(a, e)$

The transitions among these states for the baseline system is illustrated in Figure 1 from [3]. Here, the SDS asks for a piece of information and accepts it after positive verification. Otherwise, the information item is asked again. Self transitions correspond to "no reply", which is a deletion error at the front-end since the user is assumed to cooperatively reply to any query.

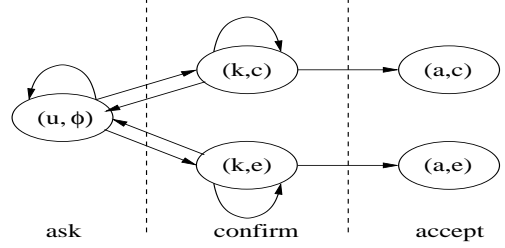


Fig. 1. Baseline finite state model (FSM0)

The transition probabilities associated with the baseline system are [3]

$$T = \begin{bmatrix} P_d & P_e & P_c & 0 & 0 \\ P_c & P_d & 0 & 0 & P_e \\ P_e & 0 & P_d & P_c & 0 \\ 0 & 0 & 0 & 1.0 & 0 \\ 0 & 0 & 0 & 0 & 1.0 \end{bmatrix} \quad (3)$$

where

$$P_e = 1 - P_c - P_d \quad (4)$$

where  $P_c$  is the probability of correct understanding, and  $P_d$  is the probability of deletion.

The second system uses the confidence estimates as

- If  $C(I|\text{ask})$  is LOW, confirm  $I$ .
- If  $C(I|\text{ask})$  is HIGH, accept  $I$ .

and

- If  $C(R|\text{confirm})$  is LOW, assume "no reply".
- If  $C(R|\text{confirm})$  is HIGH, accept  $I$ .

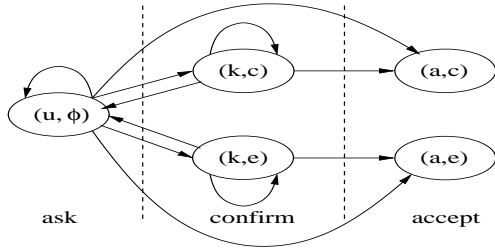
where  $I$  is the information item,  $R$  denotes "yes" or "no" reply and  $C(\cdot|\cdot)$  is the confidence score of an information item conditioned on a dialog action. The confidence score is said to be LOW if it is below a certain threshold, say  $\theta_T$ . Otherwise, it is HIGH. The corresponding finite state model is exhibited in Figure 2. The corresponding transition matrix is

$$T = \begin{bmatrix} P_d & P_e P_{CR} & P_c & P_c P_{CA} & P_e P_{FA} \\ P_c P_{CA} & P_{d'} & 0 & 0 & P_e P_{FA} \\ P_e P_{FA} & 0 & P_{d'} & P_c P_{CA} & 0 \\ 0 & 0 & 0 & 1.0 & 0 \\ 0 & 0 & 0 & 0 & 1.0 \end{bmatrix} \quad (5)$$

where

$$P_{d'} = P_d + P_e P_{CR} + P_c P_{FR} \quad (6)$$

$P_{CR}$ ,  $P_{FR}$ ,  $P_{CA}$ , and  $P_{FA}$  are the probabilities of correct rejection, false rejection, correct acceptance and false acceptance, respectively. For the details the reader is referred to [3].



**Fig. 2.** Finite state model considering confidence annotation (FSM1)

The third system has the following strategy:

- If  $C(I|\text{ask})$  is LOW, ask again.
- If  $C(I|\text{ask})$  is MEDIUM, confirm  $I$ .
- If  $C(I|\text{ask})$  is HIGH, accept  $I$ .

The confidence score is said to be MEDIUM when

$$\theta_1 < C(I|action) < \theta_2$$

This strategy has the same finite state model as in Figure 2, but the transition probabilities are different:

$$T = \begin{bmatrix} P_{d''} & P_e P_{?|e} & P_c P_{?|c} & P_c P_{CA'} & P_e P_{FA'} \\ P_c P_{CA'} & P_{d'} & 0 & 0 & P_e P_{FA'} \\ P_e P_{FA'} & 0 & P_{d''} & P_c P_{CA'} & 0 \\ 0 & 0 & 0 & 1.0 & 0 \\ 0 & 0 & 0 & 0 & 1.0 \end{bmatrix} \quad (7)$$

where

$$\begin{aligned} P_{CA'} &= 1 - P_{FR} - P_{?|c} \\ P_{FA'} &= 1 - P_{CR} - P_{?|e} \\ P_{d''} &= P_{d'} + P_e P_{?|e} + P_c P_{?|c} \end{aligned} \quad (8)$$

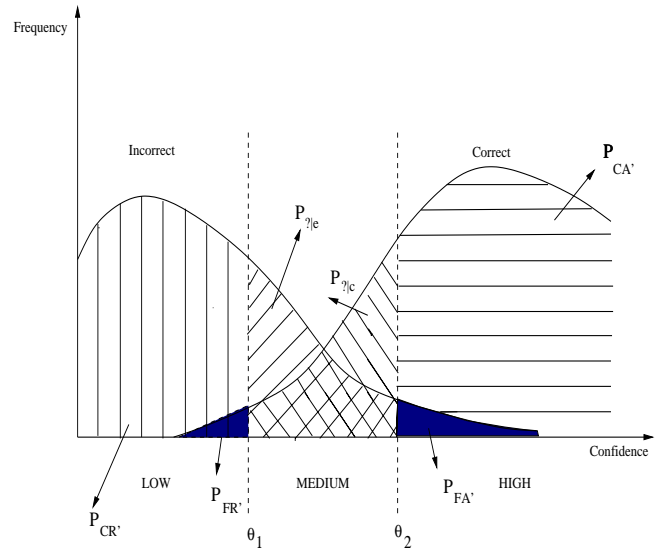
The probabilities are illustrated in Figure 3. We have selected

$$\begin{aligned} \theta_1 &= \theta_T \\ \theta_2 &= \theta_T + \Delta \end{aligned} \quad (9)$$

This allows us to have  $P_{CR}$ s and  $P_{FR}$ s the same in both systems and do fair comparison, since it is not uncommon to report the confidence estimator performance using the pair  $(P_{FR}, P_{CR})$ . Once  $\theta_T$  is fixed for the second system, one can try to see improvements (if any) with  $\Delta$  in the third system.

#### 4. NUMERICAL ANALYSIS

Numerical analysis is based on the properties of Markov chains. Typical default values for the analysis are shown in Table 1.



**Fig. 3.** Probabilities described in the text

**Table 1.** Default system parameters

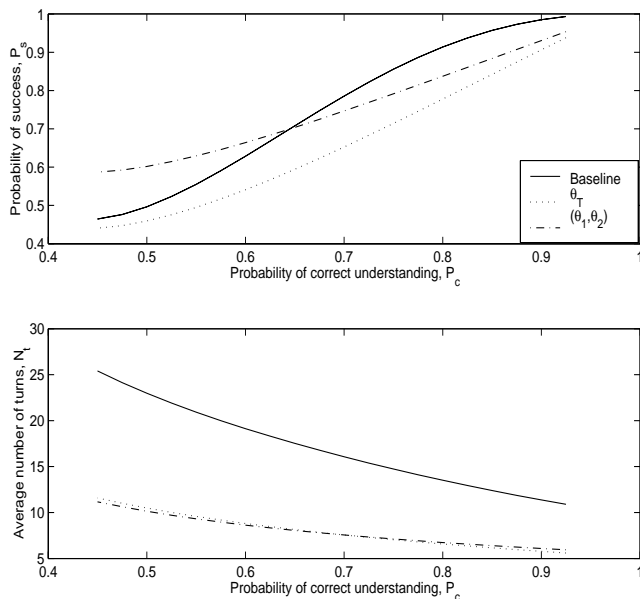
Parameter	Value
$n_i$	5
$P_{FR}$	0.05
$P_{CR}$	0.65
$P_c$	0.88
$P_d$	0.04
$P_{? c}$	0.07
$P_{? e}$	0.1

The results for all systems in terms of  $P_s$ ,  $N_t$  and  $FOM$  are presented in Table 2. By looking at  $P_s$  and  $N_t$  it is quite difficult to judge the systems. However, the  $FOM$  indicates that the SDSs with confidence are significantly better than the baseline system and slightly different from each other, which is intuitively satisfying.

We now present the behavior of SDS performance metrics with respect to  $P_c$ . Figure 4 shows  $P_s$  and  $N_t$  for all systems. An alternative to these plots with  $FOM$  is illustrated in Figure 5. One can clearly see the intuitively satisfying ordering of the systems in Figure 5.

**Table 2.** Results for example SDSs

	$P_s$	$N_t$	$FOM$
Baseline	0.96	12.3	0.71
$\theta_T$	0.85	5.8	0.81
$(\theta_1, \theta_2)$	0.89	6.4	0.82



**Fig. 4.** System performances in terms of  $P_s$  and  $N_t$  with respect to  $P_c$

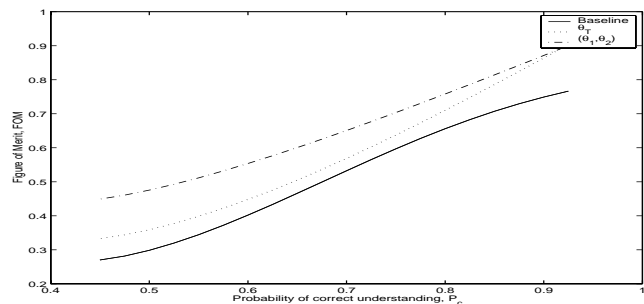
Finally we present (for the second system) the mapping of the goal  $G$  defined in terms of  $FOM$  onto the receiver operating characteristics at  $P_c = 0.92$  and  $P_d = 0.025$  in Figure 6. We also include the plot from [3] in which the goal  $G$  is defined using the component metrics. It seems that the  $G$  defined with respect to the component metrics is tighter.

## 5. CONCLUSIONS

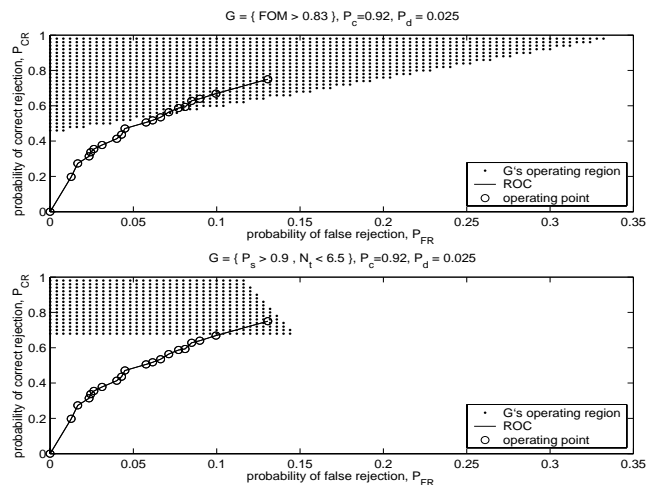
We have presented our preliminary work on developing a single metric for the quantitative analysis of spoken dialog systems. Two important metrics of a spoken dialog system, namely the success rate and efficiency, have been incorporated into a single metric. We have shown that the comparison of a number of systems using the metric has turned out to be intuitively satisfying. The metric needs to be correlated to user satisfaction by properly weighting the component metrics. Our work is in progress to determine the actual weights using the Communicator data collected at our site. In addition, we are trying to extend the analysis to more realistic scenarios by dropping some of the simplifying assumptions.

## 6. REFERENCES

[1] M. A. Walker, D. Litman, and A. Abella, “PARADISE: A framework for evaluating spoken dialog systems,” in



**Fig. 5.** System performances in terms of  $FOM$  with respect to  $P_c$



**Fig. 6.** The ROC of the confidence annotation and the operating region of the dialog goal  $G = \{FOM > 0.83\}$  (upper plot) and  $G = \{P_s > 0.9, N_t < 6.5\}$  (lower plot). The front-end performance is ( $P_c = 0.92, P_d = 0.025$ ).

*Proceedings of the 35th Annual Meeting of the Association of Computational Linguistics, 1997.*

- [2] M. A. Walker, R. Passonneau, and J.E. Boland, “Quantitative and qualitative evaluation of DARPA communicator spoken dialogue systems,” in *Meeting of the Association of Computational Linguistics*, 2001.
- [3] K. Hacioglu and W. Ward, “Does confidence annotation meet the dialog goal? A quantitative analysis,” in *Second International Conference on Human Language Technology Research*, San Diego, California,, March 24-227 2002.
- [4] B. Lin and L. Lee, “Computer-aided analysis and design for spoken dialogue systems based on quantitative simulations,” *IEEE Transactions on Speech and Audio Processing*, vol. 9, no. 5, pp. 534–548, July 2001.