Multimodal Integration—A Statistical View

Lizhong Wu, Sharon L. Oviatt, and Philip R. Cohen

***Abstract*— This paper presents a statistical approach to developing multimodal recognition systems and, in particular, to integrating the posterior probabilities of parallel input signals involved in the multimodal system. We first identify the primary factors that influence multimodal recognition performance by evaluating the multimodal recognition probabilities. We then develop two techniques, an estimate approach and a learning approach, which are designed to optimize accurate recognition during the multimodal integration process. We evaluate these methods using Quickset, a speech/gesture multimodal system, and report evaluation results based on an empirical corpus collected with Quickset. From an architectural perspective, the integration technique presented here offers enhanced robustness. It also is premised on more realistic assumptions than previous multimodal systems using semantic fusion. From a methodological standpoint, the evaluation techniques that we describe provide a valuable tool for evaluating multimodal systems.**

***Index Terms*— Combination of multiple classifiers, decision making, gesture recognition, learning, multimodal integration, speech recognition, uncertainty.**

I. Introduction

THERE are two main types of multimodal systems, one of which integrates signals at the feature level and the other at a semantic level. Systems that utilize feature fusion generally are based on multiple HMM's or temporal neural networks. In a feature fusion architecture, the correlation structure between modes can be taken into account automatically via learning. Feature fusion generally is considered more appropriate for closely coupled and synchronized modalities, such as speech and lip movements. However, such a system tends not to generalize as well if it consists of modes that differ substantially in the time scale characteristics of their features, as is the case with speech and gesture input. Modeling complexity, computational intensity, and training difficulty typically are other problems associated with the feature fusion integration approach. Due to the high dimensionality of input features and high degree of freedom of system models, a large amount of training data also is required for building this type of system. Of course, multimodal corpora rarely have been collected and labeled for training purposes, and they tend not to be publicly available so therefore are at a high premium.

Generally, multimodal systems using semantic fusion include individual recognizers and a sequential integration process. These individual recognizers can be trained using unimodal data, which are relatively easy to collect or are already publicly available for modalities like speech and handwriting. The architecture of this type of system also can leverage from existing and relatively mature unimodal recognition techniques. Such unimodal systems can be integrated directly without re-training. Compared with systems based on feature fusion, in this respect systems using semantic fusion scale up easier, whether in number of modes or size of command set.

Multimodal systems with fusion at the semantic level include Bolt's seminal work "Put-That-There" [1], ShopTalk [2], CUBRICON [3], Virtual World [4], Finger-Pointer [5], VisualMan [6], Jeanie [7], and others as described in [8]–[10]. All these previous efforts on multimodal integration have concentrated primarily on semantic representations and incorporation of new input technologies, rather than on the statistical integration process that defines a multimodal system architecture. Such systems typically also have assumed that the individual modes in a multimodal interaction function independently of each other. As a result, a multimodal command's posterior probability has been the cross product of the posterior probabilities of the associated constituents. Although this independence assumption has provided a starting point and it simplifies the integration process, it nonetheless is a naive assumption since speech and lip movements or speech and manual gestures are known to be highly correlated [11].

It also is known that a constituent in one mode typically associates with only a limited number of constituents in another mode, and that input modes differ in both their information content and recognition accuracy. An additional problem with past multimodal architectures is that the overall recognition accuracy of different input modes has been assumed to be equally reliable, although this is rarely the case. Even within the same mode, recognition accuracy varies considerably from one constituent to another. By refining the multimodal integration process so that different weights are assigned to different modes and different constituents, recognition errors potentially could be avoided so that overall system robustness is enhanced. For example, the study in [12] has shown the potential for improving continuous gesture recognition results based on a co-occurrence analysis of different gestures with

Manuscript received May 28, 1999; revised September 11, 1999. This work supported in part by DARPA under Contracts DABT63-95-C-007 and N66001-99-D-8503, and in part by ONR Grant N00014-95-1-1164. The views and conclusions contained in this paper are those of the authors and should not be interpreted as necessarily representing DARPA or ONR. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. U. Neumann.

L. Wu was with the Center for Human and Computer Communication, Department of Computer Science and Engineering, Oregon Graduate Institute of Science and Technology, Portland, OR 97291-1000 USA. He is now with HNC Software, Inc., San Diego, CA 92121–3728 USA (e-mail: Lw@hnc.com).

S. L. Oviatt and P. R. Cohen are with the Center for Human and Computer Communication, Department of Computer Science and Engineering, Oregon Graduate Institute of Science and Technology, Portland, OR 97291-1000 USA (e-mail: oviatt@cse.ogi.edu; pcohen@cse.ogi.edu).

Publisher Item Identifier S 1520-9210(99)09540-1.

spoken keywords. Performance improvement also has been found in audiovisual speech recognition systems by adaptively weighting both the audio and visual recognition channels [13]–[15].

In theory, the optimal weights for combining the posterior probabilities from different modes can be determined by the mode-conditional input feature density functions, as will be described in Section IV of this paper. In practice, it is difficult or even impossible to evaluate these conditional density functions because of the high dimensional input features. In this paper, we have developed two modeling techniques to approximate these conditional densities and to obtain the class-dependent weighting parameters for the posterior probabilities.

Another critical issue involved in the multimodal integration process is to identify the primary factors that influence multimodal recognition performance and to evaluate and estimate system recognition performance. Given a set of individual recognizers with known accuracies, is the multimodal system's performance bounded? If so, what are the theoretical lower and upper performance bounds? By estimating performance bounds, it becomes possible to evaluate the performance of alternative integration techniques in comparison with a theoretical optimum. From a diagnostic perspective, it also becomes possible to identify key factors that influence multimodal performance. To our knowledge, past research has not estimated performance bounds for guiding the development of multimodal systems.[1] In this paper, we have derived a lower and upper bound of multimodal recognition performance, and we identify the factors that influence multimodal recognition performance.

The outline of this paper is as follows. Section II briefly introduces **Quickset** [16], a multimodal system developed at OGI that has been a testbed for our research. Section III derives the multimodal recognition probability and identifies the primary factors that influence multimodal recognition performance. Section IV provides a theoretical solution for combining the multimodal posterior probabilities and discusses the problems in realizing this theoretical optimum. Sections V and VI develop two practical techniques, an estimate approach and a learning approach, designed to integrate and optimize the multimodal posterior probabilities. Empirical evaluation and performance comparisons are reported in Section VII. We summarize our findings in Section VIII.

II. Multimodal System

Our multimodal system, called **Quickset** [16], consists of parallel recognizers for the speech and gesture input modalities, which are fused at the semantic level. A command in this multimodal system is represented by constituents that are joined from the two different input modes. Each constituent is a target of an individual mode's recognizer. During the recognition process, an individual recognizer analyzes a set of input features and then produces the constituents as an N-best list of alternatives, along with posterior probabilities. **Quickset** integrates multimodal input in the following three sequential steps:

1) Temporally, **Quickset** combines speech and gesture input that is overlapped, or that falls within a certain lag relation when signals arrive sequentially. The temporal constraints of **Quickset**'s integration were determined by empirical research with users [11]. It was found that when users speak and gesture in a sequential manner, they gesture first, then speak within a relatively short time window; speech rarely precedes gesture. As a consequence, the multimodal synchronizer in **Quickset** prefers to integrate a gesture with speech that follows within a 4-s interval[2], rather than integrating it with preceding speech. If speech arrives after that interval, the gesture is interpreted unimodally. The precise lag threshold adopted when signals arrive sequentially can be learned by the system using training data, or pre-set by the system developer for a particular domain.
2) Statistically, **Quickset** integrates the posterior probabilities of constituents from individual modes, and then generates an N-best list for a multimodal command that includes posterior probabilities for each final interpretation. The original version of **Quickset** relied on the independence assumption. It took the cross product of the probabilities of individual modes to derive the multimodal probability for each item in the final multimodal N-best list. One goal of the present work is to supersede the independence assumption by developing a more powerful statistical integrator based on the realities of empirical data.
3) Semantically, **Quickset** determines whether a given gestural and spoken element in the N-best lists can be combined legally into a coherent multimodal interpretation that is executable by the system. The semantic information contained within the two modes in **Quickset** is represented as typed feature structures [17], which can be unified if the elements are compatible semantically. The unification of typed feature structures in **Quickset** has been detailed elsewhere [18].

Quickset has supported various map-based applications that enable users to set up and control distributed interactive simulations. The research and evaluations presented in this paper are based on **Quickset**'s fire/flood management corpus. Fig. 1 shows some examples of multimodal commands from this corpus.

Using a "Wizard-of-Oz" research paradigm, it was demonstrated that a multimodal interface parallel to **Quickset** supported 36% fewer task errors, 50% less disfluent input, and 10% faster task completion time than a unimodal spoken interface [19]. Further information and videotape examples of **Quickset** can be found at http://www.cse.ogi.edu/CHCC.

[1] Multimodal integration is related to, but different from previous studies on combining multiple classifiers. In multimodal integration, each individual mode recognizes a semantic constituent of commands. Different modal classifiers have different recognition targets. In combination of multiple classifiers, all classifiers share the same set of targets. This difference will be elaborated in Section III.

[2] We have found that, from about 1539 **Quickset** command patterns, more than 99% commands lie within this 4-s interval.

SPEECH	GESTURE
"burn line"	
"hot spot zone"	
"zoom out"	
"show number of gallons"	
"windspeed 40 miles per hour"	

Fig. 1. Examples of multimodal commands composed of spoken and pen-based elements in the fire/flood management corpus.

III. Primary Factors of Multimodal Recognition Performance

In a speech/gesture multimodal system, assume that $S_i, i = 1, \cdots, M$ is the output from the speech mode, and $G_j, j = 1, \cdots, N$ is the output from the gesture mode. The system is designed to recognize $C_k, k = 1, \cdots, K$ multimodal classes. The number of multimodal classes K cannot be larger than the number of attainable integrated classes $(M+1)(N+1)-1$, but will at least equal the larger number of the two output modes $\max[M, N]$.

We define the projection between the index of multimodal classes and the indices of the individual modal output as a *multimodal associative map*, as depicted by Table I. For a given corpus, the associative map defines all meaningful relations that exist between the set of speech constituents and the set of gesture constituents for each multimodal command. In the present corpus, there were 17 feature structure types for speech input (e.g., create feature object, zoom to point), eight feature structure types for pen input (e.g. line, area), and 20 feature structure types representing different types of multimodal commands (e.g., see Fig. 1). In our work, we have used the feature structure type as the basic unit for statistical integration. During multimodal recognition, the defined associative map between speech and gesture feature structure types supports a simple process of table lookups. This table can be defined directly by a user, or it can be built automatically using labeled data.

From the structure of an associative map, it is clear that the integration of this type of multimodal recognizer differs from the combination of multiple classifiers [20] or traditional data fusion [21]. The latter two are a special case of the former in which the component classifiers and the combined classifier share the same set of targets, $K = M = N$, typically displayed as a diagonal matrix.

The multimodal system can avoid some recognition errors that otherwise would occur in a unimodal system simply by checking whether recognized speech and gesture pieces can be integrated legally or not, given the system's semantic constraints. This type of error avoidance occurs as long as the

TABLE I
Multimodal Associative Map for the Fire/Flood Management Corpus, Representing the Complete Set of Legitimate Semantic Combinations Possible Between All Types of Spoken and Pen-Based Input for the 20 Multimodal Feature Types in this Corpus. For Example, the First Type Represents a Combination Between the First Speech Class and the Fourth Gesture Class

MULTI		GESTURE CLASS							
MODAL		1	2	3	4	5	6	7	8
	1	0	0	0	1	0	0	0	2
S	2	0	0	0	3	0	0	0	0
P	3	4	0	0	0	0	0	0	0
E	4	0	0	0	5	0	0	0	0
E	5	0	6	7	0	0	0	0	0
C	6	8	0	0	0	0	0	0	0
H	7	0	0	0	0	9	0	0	0
	8	0	0	0	0	0	10	0	0
	9	0	11	12	0	0	0	0	0
	10	0	0	0	0	0	0	13	0
C	11	0	0	0	0	0	14	0	0
L	12	0	0	0	0	0	15	0	0
A	13	0	0	0	0	0	0	0	16
S	14	0	0	0	17	0	0	0	0
S	15	0	0	0	0	18	0	0	0
	16	19	0	0	0	0	0	0	0
	17	20	0	0	0	0	0	0	0

number of multimodal classes is less than the maximum number of potentially attainable multimodal classes, i.e. $K < (M+1)(N+1)-1$. Evidence for the error compensation that results during unification of typed feature structures has been detailed previously [22]. In the following, we discuss multimodal performance and establish its bound from a statistical point of view.

Assume that X represents a multimodal input feature vector, which is a combination of gesture input feature X^G and and speech input feature X^S. A system designed to recognize K classes of commands will partition the input feature space into K disjoint decision regions $R_k, k = 1, \cdots, K$. The probability of correct recognition for X is thus

$$P_c = \sum_{k=1}^{K} P(X \in R_k, C_k) = \sum_{k=1}^{K} P_{X_k} P(C_k) \qquad (1)$$

where

$$P_{X_k} = P(X \in R_k | C_k) = \int_{R_k} p(X|C_k)\, dX \tag{2}$$

is the probability of correct recognition for the kth class, $p(X|C_k)$ is its class-conditional density and $P(C_k)$ is its prior probability.

By expressing the multimodal recognition probability using the recognition probabilities of its associated modes, we have obtained the following multimodal recognition probability bound (see Appendix A):

$$\sum_{k=1}^{K} P_{X_j^G} P_{X_i^S} P(C_k) \le P_c \le \sum_{k=1}^{K} \max[P_{X_j^G}, P_{X_i^S}] P(C_k) \tag{3}$$

where $P_{X_j^G}$ and $P_{X_i^S}$ are respectively the correct recognition probabilities of the associated gesture, and speech constituents for the kth multimodal class.

Equation (3) reveals that a multimodal system performs at its lower bound if individual modes are assumed to be independent, and a simple joint probability estimate is calculated during integration. In contrast, a multimodal system performs at its upper bound when the information in one mode is completely redundant with that in the other mode. In general, we would summarize that multimodal recognition performance is determined by the following factors:

1) recognition accuracy of the individual modes;
2) structure of the *associative map*;
3) manner of combining posterior probabilities;
4) prior distribution of multimodal commands.

IV. Integration of Multimodal Posterior Probabilities

During recognition, our goal is to evaluate the posterior probabilities of all multimodal classes, given an unknown input feature set. Combining the posterior probabilities involves combining the class-conditional density function $p(X|C_k)$. As shown in Appendix B, we have found that

$$p(X|C_k) = \tfrac{1}{2}[p(X^S|C_k, X^G)p(X^G|C_k) + p(X^G|C_k, X^S)p(X^S|C_k)] \tag{4}$$

where $p(X^S|C_k)$ and $p(X^G|C_k)$ are the class-conditional densities estimated by the speech recognizer and gesture recognizer, and $p(X^G|C_k, X^S)$ and $p(X^S|C_k, X^G)$ are the mode-conditional input feature densities for the kth class.

Equation (4) provides a theoretical solution for integrating multimodal class-conditional density functions. However, due to the high dimensional input features, a large amount of training data is required to evaluate the mode-conditional input feature densities directly. A conventional input representation in acoustic modeling uses a 39-dimensional vector (i.e., the signal energy and first 12 cepstral coefficients and their first- and second-order differentials [23]) for each 10-ms speech block. This means that the speech input dimension will increase to 3900, even if a voice command lasts for only 1 s. An example of pen input feature representations is Apple Computer's Newton handwriting recognizer, in which the input dimension is 382 (i.e., 14×14 image, 20×9 stroke features, 5-dimensional stroke count, and single-dimension aspect ratio [24]). Based on these examples, a multimodal input feature dimension could easily be as large as 4282. Considering the "curse of dimensionality" in data modeling[3], if the data for a given sampling density in one dimension total 30, then the total required for multimodal input feature modeling would be 30^{4282}.

The above calculations reveal why it is hard to obtain an estimate of mode-conditional input feature density functions. Accurate estimates also are difficult to obtain because few actual multimodal corpora are available. Therefore, evaluation of (4) requires approximation. By letting

$$\alpha_k = \tfrac{1}{2} p(X^S|C_k, X^G) \tag{5}$$

$$\beta_k = \tfrac{1}{2} p(X^G|C_k, X^S) \tag{6}$$

(4) can be rewritten as

$$p(X|C_k) = \alpha_k p(X^G|C_k) + \beta_k p(X^S|C_k). \tag{7}$$

α_k and β_k become the weighting parameters to the modal class-conditional densities. They are still class-conditional, but independent of individual input features. In the next two sections, we develop two techniques to evaluate these parameters.

V. Estimate Approach

Here, the normalized mode-conditional recognition probabilities[4] are taken as an approximation of the mode-conditional input feature densities. That is

$$\alpha_k \approx \frac{1}{P_0} P(X^S \in R_i^S | C_k, X^G \in R_j^G) \tag{8}$$

$$\beta_k \approx \frac{1}{P_0} P(X^G \in R_j^G | C_k, X^S \in R_i^S) \tag{9}$$

where P_0 is a normalization factor and

$$P_0 = P(X^S \in R_i^S | C_k, X^G \in R_j^G) + P(X^G \in R_j^G | C_k, X^S \in R_i^S). \tag{10}$$

It is much easier to evaluate the mode-conditional recognition probabilities than to evaluate the mode-conditional input feature density functions. Therefore, two methods are developed to estimate the mode-conditional recognition probabilities, with the preferred method depending on the availability of training data. Method-I estimates the conditional probabilities by simply counting the number of nonzero entries in each column and row of the *associative map*. For example, for the first multimodal class shown in Table I, which associates with the first speech class and the fourth gesture class, there are two nonzero entries in the corresponding row and four

[3] The curse of dimensionality [25], [26] refers to the exponential growth of hypervolume as a function of dimensionality. If N is the total data for a given sampling density in one dimension, then when the dimensionality is increased to m the total data must also increase to N^m to keep the same sampling density.

[4] A mode-conditional recognition probability is a conditional probability for recognizing one mode, given information about the other. Further description is available in Appendix A.

nonzero entries in the corresponding column, or $\hat{P}(X^S \in R_1^S | C_1, X^G \in R_4^G) = \frac{1}{4}$ and $\hat{P}(X^G \in R_4^G | C_1, X^S \in R_1^S) = \frac{1}{2}$. After normalization, $\hat{\alpha}_1 = 0.33$ and $\hat{\beta}_1 = 0.67$. The advantage of this particular method is that it does not require any training data.

When training data are available, Method-II provides a more accurate estimate of the mode-conditional recognition probabilities. Method-II is a bin-counting process. For the kth multimodal class, the patterns labeled as the kth multimodal class in the training data are located first. Among these multimodal patterns, it is assumed that there are N_k^G patterns having correct gesture output, N_k^S patterns having correct speech output, and N_k patterns having both correct gesture and speech output. The estimates of mode-conditional recognition probabilities then are

$$\hat{P}(X^S \in R_i^S | C_k, X^G \in R_j^G) = \frac{N_k}{N_k^G} \tag{11}$$

$$\hat{P}(X^G \in R_j^G | C_k, X^S \in R_i^S) = \frac{N_k}{N_k^S} \tag{12}$$

and the estimates of the weighting parameters are

$$\hat{\alpha}_k = \frac{N_k^S}{N_k^G + N_k^S} \tag{13}$$

$$\hat{\beta}_k = \frac{N_k^G}{N_k^G + N_k^S}. \tag{14}$$

From (13) and (14), it is clear that the weighting parameters α and β depend only on the ratio of recognition rates of the individual modes. If the ratio between the gesture and speech recognition probabilities for the kth multimodal class is defined as

$$\lambda_k = \frac{P_{X_j^G}}{P_{X_i^S}} = \frac{N_k^G}{N_k^S}, \tag{15}$$

then

$$\hat{\alpha}_k = \frac{1}{1 + \lambda_k} \tag{16}$$

$$\hat{\beta}_k = \frac{\lambda_k}{1 + \lambda_k}. \tag{17}$$

As shown, if both modes perform equally well and the ratio is about 1, then both modes will be equally weighted. If one mode's output is significantly biased toward low performance, then it will be given a larger weight to correct this bias.

VI. Members to Teams to Committee: The MTC Approach

MTC is a novel recognition technique developed to build a complex pattern recognition system with high-dimensional input features [27]. In this section, we first provide an introduction of the MTC technique by presenting its overall architecture, the functionalities of each component, and its learning algorithms. We then describe the application of the MTC approach to integrating a multimodal system.

A. MTC Architecture

The MTC architecture consists of three layers. The bottom layer is formed by multiple recognizer members. Each member is a local posterior estimator with an assigned input variable subset, a specified model type and complexity, and a given training and validation data set. The members cooperate with each other via the multiple teams built at the mid-layer. Different teams observe different training data, and are initialized and trained differently. The team integrates the members. Multiple teams are built to reduce integration uncertainty. Output from the teams forms an empirical posterior distribution that then is sent to the committee at the upper layer. The committee makes a final decision after comparing the empirical posterior distributions of different targets.

B. MTC Recognition Algorithm

In general, we define the input feature set $I = \{I_1, I_2, \cdots, I_n\}$ and the recognition target set $T = \{T_1, T_2, \cdots, T_m\}$. The input feature I is formed by n-streams, whose dimensions may differ. The target T consists of m different classes, for example of different multimodal commands. The MTC recognition algorithm goes through three bottom-up steps.

1) Estimating the local posteriors of members: Each member computes a local posterior estimate under the specified modeling condition. The modeling specifications include the model type, the model complexity, the extraction of input features, the training and validation data, and the learning algorithm. If there is a total of M combinations of modeling specifications in which we are interested, then we would compute M local posterior estimates from the M members as follows:

$$\hat{P}(T_k | I, S_i), \quad \text{with } k = 1, \cdots, m \quad \text{and} \quad i = 1, \cdots, M \tag{18}$$

where S_i stands for the ith combination of modeling specifications.

2) Coordinating the local posteriors into teams: The team integrates the local posterior estimates of different specifications. We have

$$\hat{P}(T_k | I) = \sum_{i=1}^{M} \hat{P}(T_k | I, S_i) P_k(S_i), \quad \text{for } k = 1, \cdots, m \tag{19}$$

where $P_k(S_i)$ is the mode probability of the kth target associated by the ith combination of modeling specifications. The team is trained to learn the mode probability matrix. Different training data and approaches will result in different mode probability estimates. Subsequently, the multiple team posteriors are obtained:

$$\hat{P}^{(l)}(T_k | I), \quad \text{with } k = 1, \cdots, m \quad \text{and} \quad l = 1, \cdots, L \tag{20}$$

where l is the index of ways of estimating the mode probability and L is the total number of ways that we are interested in.

3) Making a recognition decision via committee: The output from multiple teams forms an empirical distribution

of posterior $P(T_k|I)$, which is approximated by a normal or t-student distribution, depending on the size of the samples. Given a confidence level, the committee runs through a series of pair-by-pair hypothesis tests and obtains a significance matrix H. H is an $m \times m$ square matrix, where m is the number of recognition targets. The element of H, h_{ij}, is either 1 when the posterior estimate $\{\hat{P}^{(l)}(T_i|I), l = 1, \cdots, L\}$ is significantly greater than $\{\hat{P}^{(l)}(T_j|I), l = 1, \cdots, L\}$ within the given confidence level, -1 when it is significantly less, or otherwise 0. By definition, the diagonal elements of H are all zeros, if $h_{ij} = 1$ then $h_{ji} = -1$; if $h_{ij} = 0$ then $h_{ji} = 0$; and $\Sigma_i \ \Sigma_j \ h_{ij} = 0$. The recognition targets then are ranked by summing over each row in H, and this summary value is called a *significance number*. All significance numbers form an m-dimensional significance vector V. The maximal significance number is $m - 1$. If there is a significance number that equals $m-1$, the input is recognized as the target corresponding to the row index of this maximal significance number. If all significance numbers are smaller than $m - 1$, then the current input cannot be recognized with confidence, and further external information is required.

C. MTC Training Procedure

The goal of the first tier involving the MTC's members is to learn a set of local posterior estimates, as indicated in (18). The key of this first layer is to identify the modeling specifications and their combinations. The members within the MTC can represent different types of models, with the training algorithm for the members being model-dependent. Among the various modeling specifications, the most important one is the extraction of input features via exploratory data analyzes. Once the input features have been extracted, the model type is selected to fit the characteristics of these input features. In the MTC, a variety of input features can be extracted, and different types of models can be selected to fit different types of input features.

With respect to the training procedure for teams, the goal of this second layer is to learn the mode probability matrix in (19). By adjusting the mode probability matrix $\{P_k(S_i)\}$, we maximize (i.e., reward) the kth posterior $\hat{P}(T_k|I)$ and simultaneously reduce (i.e., penalize) the other posteriors $\hat{P}(T_j|I)$, for $j = 1, \cdots, m$ and $j \neq k$, when the kth-class pattern is applied. In order to meet the constraint that the sum of all posteriors must equal one, we impose a softmax function [28] on the output. The detailed learning algorithm is given in [27].

The team integrates the members' posterior estimates. Multiple teams are built to reduce integration uncertainty. The goal of the committee is to compare the empirical posterior distribution formed by the teams and make a final recognition decision. To train the committee, no free system parameter is needed. The confidence level for recognition is predetermined, and different confidence levels will result in a different system error rate/rejection rate tradeoff. The higher the confidence level, the lower the error rate but the higher the rejection rate.

TABLE II
PERCENT CORRECT COMMAND RECOGNITION RATES FOR TEST DATA REPRESENTING THE DIFFERENT POSTERIOR PROBABILITY INTEGRATION TECHNIQUES DESCRIBED IN SECTIONS V AND VI, AS WELL AS THE UPPER AND LOWER PERFORMANCE BOUNDS DESCRIBED IN SECTION III

Lower Bound	Estimate I	Estimate II	MTC	Upper Bound
78.91	87.78	88.15	95.26	96.65

D. MTC Multimodal Statistical Integration

Our proposed MTC technique is well suited to experimenting with ways to integrate multiple modes on the basis of posterior probabilities and other factors. Using this technique, the recognizers of different modes become the members of an MTC statistical integrator. Multiple teams built in the MTC integrator are trained to coordinate and weight the output from different modes. Each team establishes the posterior estimate for a multimodal command, given the current multimodal input received. The committee of the MTC integrator analyzes the empirical distribution of the posteriors and establishes the N-best ranking for each multimodal command.

VII. EMPIRICAL RESULTS

Our **Quickset** system was the testbed used to formulate and evaluate the derived performance bounds and the proposed integration concepts. As mentioned earlier, the data corpus used in the present work was collected using **Quickset** while users performed community fire and flood management tasks. All commands were multimodal, involving both speech and gesture input. This corpus consisted of 1539 labeled commands collected from sixteen users, eight native speakers of English and eight accented nonnative speakers. We randomly assigned the data from the first eight users for development, and the rest for test purposes. As illustrated in Table I, there were 17 feature structure types for speech, eight for pen, and 20 representing all types of multimodal commands. With this arrangement, each of the 20 basic units had an average of 40 training patterns. More detailed description and data analyzes on this corpus have been described elsewhere [22].

Table II summarizes our empirical evaluation based on this corpus. The recognition performance ranged from 78.91% correct at the lower bound, to 96.65% correct at the upper bound. The columns Estimate I & II correspond to Method I & II of the estimate approach described in Section V, and the MTC column corresponds to the learning approach described in Section VI. As expected, all performance lies within the theoretical lower and upper bounds. Estimate II performs only slightly better than Estimate I, and both are significantly worse than the MTC approach—which only departs 1.4% from the system's established theoretical optimum.

The favorable performance of the MTC approach can be attributed to several factors. First, the MTC approach adopted a discriminative training scheme, which maximizes or rewards the correct class-conditional density and simultaneously reduces or penalizes others. Secondly, with the MTC approach, multiple sets of weighting parameters were trained, with each

set providing an estimate of the class-conditional density function. A committee of multiple sets provides a smoother estimate function than any individual one, which leads to better robustness and generalization. Thirdly, the MTC approach takes into account the fact that the class-conditional densities of individual modalities may be normalized differently, which is a pragmatic reality when the recognizers are developed by different sources. Through training that enables the learning of weighting parameters, the MTC approach is able to normalize the output from different recognizers to the same scale.

VIII. Conclusions

The development of an architecture for integrating different input modes and for evaluating multimodal recognition performance are two critical issues for the development of next-generation multimodal systems. In this paper, we have evaluated the multimodal recognition probabilities. It was revealed that the multimodal recognition performance, in general, is determined by the recognition accuracy of individual modes, the structure of the *associative map*, the manner of combining posterior probabilities, and the prior distribution of multimodal classes.

In theory, the optimal weights for combining multimodal posterior probabilities can be determined by the mode-conditional input feature density functions. In practice, it is difficult or even impossible to evaluate these conditional density functions because of the high dimensional input features. Therefore, we have developed two techniques to approximate these conditional densities, and obtained the class-dependent weighting parameters for the posterior probabilities. The first technique is an estimate approach in which the mode-conditional input feature density is approximated by the normalized mode-conditional recognition probability. The latter then can be estimated based on the structure of the *associative map*, or using the labeled training data. The second technique is a learning approach in which the weighting parameters are trained to maximize the correct posterior probability and minimize the wrong ones. Several key learning techniques also have been incorporated into this mechanism to improve the robustness and generalization of its performance.

The integration techniques and evaluation tools presented in this paper provide a statistical approach to developing multimodal recognition systems. We have evaluated these new methods using **Quickset** and an empirical corpus collected with **Quickset**. Although the current version of **Quickset** is a speech/gesture bimodal system, our proposed techniques offer a general architectural approach that could be extended to multimodal systems involving other modes or more than two modes.

Appendix

A. Derivation of (3)

By expressing the joint probability in conditional probabilities, (2) can be re-written as

$$P_{X_k} = P(X^G \in R_j^G, X^S \in R_i^S | C_k) \tag{21}$$

$$= P(X^G \in R_j^G | C_k, X^S \in R_i^S) P_{X_i^S} \tag{22}$$

$$= P(X^S \in R_i^S | C_k, X^G \in R_j^G) P_{X_j^G} \tag{23}$$

where R_j^G and R_i^S are the partitioned decision rejoins in the gestural feature space and the spoken feature space. They are associated to the kth multimodal class. $P_{X_j^G}$ and $P_{X_i^S}$ are respectively the correct recognition probabilities of the gesture recognizer and the speech recognizer for the kth multimodal class. $P(X^G \in R_j^G | C_k, X^S \in R_i^S)$ and $P(X^S \in R_i^S | C_k, X^G \in R_j^G)$ are the conditional probabilities for recognizing one mode, given information on the other.

Since the conditional probability will not be less than the probability with the condition being removed, from (22) or (23), we have

$$P_{X_k} \geq P_{X_j^G} P_{X_i^S}. \tag{24}$$

It is clear that the lower bound will be obtained when the gesture and speech recognition modes are independent of each other.

Since any probability is upper bounded by one, the upper bound of P_{X_k} is

$$P_{X_k} \leq \max[P_{X_j^G}, P_{X_i^S}]. \tag{25}$$

This upper bound will be obtained when one mode is completely redundant with another.

Substituting the above inequalities (24) and (25) into (1), we obtain (3).

B. Derivation of (4)

Analogously to (22) and (23), we have

$$p(X|C_k) = p(X^G, X^S | C_k) \tag{26}$$

$$= p(X^G | C_k, X^S) p(X^S | C_k) \tag{27}$$

$$= p(X^S | C_k, X^G) p(X^G | C_k) \tag{28}$$

Summing over (27) and (28), we obtain (4).

Acknowledgment

The authors would like to thank D. McGee for his valuable comments, as well as other members of the CHCC at OGI for their suggestions. We would also like to thank the reviewers for their comments.

References

[1] R. A. Bolt, "Put that there: Voice and gesture at the graphics interface," *Comput. Graph.*, vol. 14, no. 3, pp. 262–270, 1980.

[2] P. R. Cohen, M. Dalrymple, D. B. Moran, F. C. N. Pereira, J. W. Sullivan, R. A. Gargan, J. L. Schlossberg, and S. W. Tyler, "Synergistic use of direct manipulation and natural language," in *Human Factors in Computing Systems: CHI'89 Conference Proceedings.* New York: Addison-Wesley, Apr. 1989, pp. 227–234; reprinted in *Readings in Intelligent User Interfaces*, M. Maybury and W. Wahlster, Eds. San Francisco, CA: Morgan Kaufman, 1998.

[3] J. G. Neal and S. C. Shapiro, "Intelligent multimedia interface technology," in *Intelligent User Interfaces*, J. Sullivan and S. Tyler, Eds. New York: ACM, 1991, pp. 11–43.

[4] C. Codella, R. Jalili, L. Koved, J. Lewis, D. Ling, J. Lipscomb, D. Rabenhorst, C. Wang, A. Norton, P. Sweeney, and C. Turk, "Interactive simulation in a multi-person virtual world," in *Proc. ACM Conf. Human Factors in Computing Systems (CHI'92)*, 1992, pp. 329–334.

[5] M. Fukumoto, Y. Suenaga, and K. Mase, "Finger-pointer: pointing interface by image processing," *Comput. Graph.*, vol. 18, no. 5, pp. 633–642, 1994.

[6] J. Wang, "Integration of eye-gaze, voice and manual response in multimodal user interface," in *Proc. IEEE Int. Conf. Systems, Man and Cybernetics*, 1995, pp. 3938–3942.

[7] M. T. Vo and C. Wood, "Building an application framework for speech and pen input integration in multimodal learning interfaces," in *Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing*, Atlanta, GA, 1996, pp. 3545–3548.

[8] D. B. Koons, C. J. Sparrell, and K. R. Thorisson, "Integrating simultaneous input from speech, gaze and hand gestures," in *Intelligent Multimedia Interfaces*, M. Maybury, Ed. Cambrudge, MA: MIT Press, 1993, pp. 257–276.

[9] R. Sharma, T. S. Huang, V. I. Pavlović, Y. Zhao, Z. Lo, S. Chu, K. Schulten, A. Dalke, J. Phillips, M. Zeller, and W. Humphrey, "Speech/gesture interface to a visual computing environment for molecular biologists," in *Proc. Int. Conf. Pattern Recognition*, Aug. 1996, pp. 964–968.

[10] R. Sharma, V. I. Pavlović, and T. S. Huang, "Toward multimodal human-computer interface," *Proc. IEEE, Special Issue on Multimedia Signal Processing*, vol. 86, pp. 853–869, May 1998.

[11] S. Oviatt, A. DeAngeli, and K. Kuhn, "Integration and synchronization of input modes during multimodal human-computer interaction," in *Proc. Conf. Human Factors in Computing Systems: CHI'97*, 1997, pp. 415–422.

[12] I. Poddar, Y. Sethi, E. Ozyildiz, and R. Sharma, "Toward natural gesture/speech HCI: A case study of weather narration," in *Proc. 1998 Workshop on Perceptual User Interfaces (PUI'98)*, M. Turk, Ed., San Francisco, CA, Nov. 1998, pp. 1–6.

[13] U. Meier, W. Hurst, and P. Duchnowski, "Adaptive bimodal sensor fusion for automatic speechreading," in *Proc. Int. Conf. Acoustics, Speech, and Signal Processing*, 1996, pp. 833–836.

[14] T. Sejnowski, B. Yuhas, M. Goldstein, and R. Jenkins, "Combining visual and acoustic speech signal with a neural network improves intelligibility," in *Advances in Neural Information Processing Systems*, D. Touretzky, Ed. San Francisco, CA: Morgan Kaufman, 1990, pp. 232–239.

[15] A. Waibel, M. T. Vo, P. Duchnowski, and S. Manke, "Multimodal interfaces," *Artif. Intell. Rev., Special Volume on Integration of Natural Language and Vision Processing*, vol. 10, no. 3-4, pp. 299–319, Aug. 1995.

[16] P. Cohen, M. Johnston, D. McGee, S. Oviatt, J. Pittman, I. Smith, L. Chen, and J. Clow, "Quickset: Multimodal interaction for distributed applications," in *Proc. Fifth ACM Int. Multimedia Conf.*, New York, 1997, pp. 31–40.

[17] R. Carpenter, *The Logic of Typed Feature Structures*. Cambridge, U.K., Cambridge Univ. Press, 1992.

[18] M. Johnston, P. Cohen, D. McGee, S. Oviatt, J. Pittman, and I. Smith, "Unification-based multimodal integration," in *Proc. 35th Annu. Meeting Association for Computational Linguistics*, San Francisco, CA, 1997, pp. 281–288.

[19] S. Oviatt, "Multimodal interfaces for dynamic interactive maps," in *Proc. Conf. Human Factors in Computing Systems: CHI'96*, Vancouver, B.C., Canada, 1996, pp. 95–102.

[20] J. Kittler, M. Hatef, R. Duin, and J. Matas, "On combining classifiers," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 20, pp. 226–239, Mar. 1998.

[21] M. A. Abidi and R. C. Gonzalez, *Data Fusion in Robotics and Machine Intelligence*. Boston, MA: Academic, 1992.

[22] S. Oviatt, "Mutual disambiguation of recognition errors in a multimodal architecture," in *Proc. Conf. Human Factors in Computing Systems: CHI'99*, Pittsburgh, PA, 1999, pp. 576–583.

[23] S. Young, "Large vocabulary continuous speech recognition: A review," Tech. Rep., Engineering Dept., Cambridge Univ. Cambridge, U.K., 1996.

[24] L. S. Yaeger, B. J. Webb, and R. F. Lyon, "Combining neural networks and context-driven search for online, printed handwriting recognition in the Newton," *AI Mag.*, vol. 19, no. 1, pp. 73–89, Spring 1998.

[25] R. O. Duda and P. E. Hart, *Pattern Classification and Scene Analysis*. New York: Wiley, 1973.

[26] J. H. Friedman, "An overview of predictive learning and function approximation," in *From Statistics to Neural Networks, Theory and Pattern Recognition Applications*, V. Cherkassky and J. H. Friedman, Eds., pp. 1–61. NATO ASI Series F, vol. 136. Berlin, Germany: Springer, 1994.

[27] L. Wu, S. Oviatt, and P. Cohen, "From members to teams to committee—A robust approach to gestural and multimodal recognition," submitted for publication.

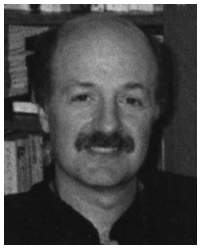
[28] J. S. Bridle, "Probabilistic interpretation of feedforward classification network outputs, with relationships to statistical pattern recognition," in *Neurocomputing: Algorithms, Architectures and Applications*, F. Fogelman Soulié and J. Hérault, Eds., pp. 227–236. New York: Springer-Verlag, 1990.

Lizhong Wu received the B.Sc. and M.Sc. degrees in electrical engineering from South China University of Science and Technology in 1983 and 1986, respectively, and his Ph.D degree in information engineering from Cambridge University, Cambridge, U.K., in 1992.

He was a Post-doctoral Fellow at Cambridge University from 1992 to 1993, a Senior Research Associate at the Oregon Graduate Institute of Science and Technology (OGI), Portland, from 1994 to 1995, and a Senior Rresearch Scientist at Nonlinear Prediction Systems, Beaverton, OR, from 1996 to 1997. He is now a Principal Project Scientist at the Center for Human and Computer Communication, OGI. His research interests include machine learning, multimedia signal processing, multimodal speech and gesture recognition.

Sharon L. Oviatt received the B.A. degree (with highest honors) from Oberlin College, Oberlin, OH, and the Ph.D. from the University of Toronto, Toronto, Ont., Canada, in 1979.

She is a Professor and Co-Director of the Center for Human-Computer Communication (CHCC) in the Department of Computer Science, Oregon Graduate Institute of Science and Technology (OGI), Portland. She previously has taught and conducted research at the Artificial Intelligence Center at SRI International, and the University of Illinois, Urbana-Champaign, University of California, Berkeley, and Oregon State University, Corvallis. Her current research focuses on human–computer interaction, interface design for multimodal/multimedia systems and speech systems, portable and telecommunication devices, and highly interactive systems. She is an active member of the international HCI and speech communities, has published over 60 scientific articles, and has served on numerous government advisory panels and editorial boards. Her work is featured in recent special issues on Multimodal Interfaces appearing in the IEEE TRANSACTIONS ON MULTIMEDIA, *Human-Computer Interaction*, and *Communications of the ACM*.



Philip R. Cohen received the B.A. degree in mathematics from Cornell University, Ithaca, NY, and the M.Sc. and Ph.D. degrees in computer science from the University of Toronto, Toronto, Ont., Canda.

He has been a Researcher or Faculty Member at Bolt Beranek and Newman, Inc., Oregon State University, Corvallis, the University of Illinois, urbana-Champaign, Fairchild Laboratory for Artificial Intelligence Research, and SRI International. He is currently a Professor and Co-director of the Center for Human-Computer Communication, Department of Computer Science, Oregon Graduate Institute of Science and Technology, Portland. His research interests include multimodal interaction, multiagent systems, dialogue, natural language processing, and theories of collaboration and communication.

Dr. Cohen is currently the President of the Association for Computational Linguistics and is a Fellow of the American Association for Artificial Intelligence.