

Cross-Modality Semantic Integration With Hypothesis Rescoring for Robust Interpretation of Multimodal User Interactions

Pui-Yu Hui, *Graduate Student Member, IEEE*, and Helen M. Meng, *Member, IEEE*

Abstract—We develop a framework pertaining to automatic semantic interpretation of multimodal user interactions using speech and pen gestures. The two input modalities abstract the user's intended message differently into input events, e.g., key terms/phrases in speech or different types of gestures in the pen modality. The proposed framework begins by generating partial interpretations for each input event as a ranked list of hypothesized semantics. We devise a *cross-modality semantic integration procedure* to align the pair of hypothesis lists between every speech input event and every pen input event in a multimodal expression. This is achieved by the Viterbi alignment algorithm that enforces the temporal ordering of the input events as well as the semantic compatibility of aligned events. The alignment enables generation of a *unimodal, verbalized paraphrase* that is semantically equivalent to the original multimodal expression. Our experiments are based on a multimodal corpus in the domain of city navigation. Application of the cross-modality integration procedure to near-perfect (manual) transcripts of the speech and pen modalities show that correct unimodal paraphrases are generated for over 97% of the training and test sets. However, if we replace with automatic speech and pen recognition transcripts, the performance drops to 53.7% and 54.8% for the training and test sets, respectively. In order to address this issue, we devised the *hypothesis rescoring procedure* that evaluates all candidates of cross-modality integration derived from multiple recognition hypotheses from each modality. The rescoring function incorporates the integration score, N -best purity of recognized spoken locative expressions, as well as distances between coordinates of recognized pen gestures and their interpreted icons on the map. Application of *cross-modality hypothesis rescoring* improved the performance to 67.5% and 69.9% for the training and test sets, respectively.

Index Terms—Joint integration, human-computer interaction, hypothesis rescoring, multimodal input, pen gesture, perplexity, robust interpretation, spoken input.

I. INTRODUCTION

WE develop a framework pertaining to automatic semantic interpretation of multimodal user interactions using speech and pen gestures. These two input modalities are gaining increasing importance in our information society,

Manuscript received February 01, 2008; revised November 24, 2008. Current version published February 11, 2009. This work was supported in part by a grant from the HK SAR Government Research Grants Council under Project CUHK4151/05E. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Timothy J. Hazen.

The authors are with Human-Computer Communications Laboratory, Department of Systems Engineering and Engineering Management, The Chinese University of Hong Kong, Shatin, N.T., Hong Kong, China (e-mail: pyhui@se.cuhk.edu.hk; hmmeng@se.cuhk.edu.hk).

Digital Object Identifier 10.1109/TASL.2008.2011509

along with rapid growth in the penetration of handheld mobile devices. The coordinated use of speech and pen gestures offers ease in direct retrieval and manipulation of information. As discussed in [1], users tend to migrate from unimodal to multimodal interactions when tackling tasks with increasing difficulty and communicative complexity. Coordinated use of both modalities enhances expressive power, especially in the communication of complex semantics in succinct form [2]. For example, the unimodal spoken inquiry:

What is the name of the street that is five blocks south of the Yonghegong and lies to the east of the China National Museum of Fine Arts?

may be paraphrased multimodally with substantial simplification, to become:

What street is this? <draw a stroke on the map>

Each modality in the multimodal user input presents a different abstraction of the user's informational or communicative goal as one or more input events. An input event may be a spoken deictic term/phrase or a pen action. The semantics of an input event may be imprecise (e.g., a pen stroke on a map may denote a street or demarcation), incomplete (e.g., use of anaphora in "how about the previous one"?), or erroneous due to misrecognitions (e.g., speech or pen gesture recognition errors). These problems motivate us to investigate 1) how we may characterize individual input events in a multimodal input expression to derive their possible (incomplete) semantics, 2) how we may combine such partial semantics across modalities to derive the holistic semantic meaning of the original multimodal expression, and 3) how we may leverage the mutual reinforcements and mutual disambiguation across modalities [3] to achieve robustness towards misrecognitions and imperfectly captured inputs.

Previous approaches towards semantic interpretation of multimodal input include frame-based heuristic integration, unification parsing, hybrid symbolic-statistical approach, weighted finite-state transducers, probabilistic graph matching and the salience-driven approach. Frame-based heuristic integration [4], [5] uses an attribute-value data structure that incorporates temporal difference and contextual information for semantic integration with a set of rules. Unification parsing [6], [7] combines temporally and semantically compatible speech/gesture recognition hypotheses that are represented as typed feature structures with multimodal grammars rules. Hybrid symbolic-statistical approach [8], [9] aims to statistically refine unification-

based parsing with probabilities and confidence scoring of the features structures in order to account for correlations between modalities. Weighted finite-state transducers (FSTs) [10] encode syntactic and semantic information to offer tight coupling across modalities, with FST weights as trained from data. Probabilistic graph matching [11], [12] incorporates semantic, temporal, and contextual constraints to combine information from multiple input modalities, where the information is represented as attribute relational graphs (ARGs). Integration includes maximizing the node match probabilities between ARG from speech and the ARG from pen input. The salience-driven approach [13] is an n -gram language model that incorporates a salience distribution based on the pen gesture to constraint the bigram probability for spoken language understanding.

We aim to devise a cross-modality (speech and pen) semantic integration framework that draws from previous experiences but is extended with several desirable features.

- 1) Ability to leverage the multiple recognition hypotheses generated from speech and pen recognition—we devise a score-based integration process that considers the ranked confidence of multiple recognition hypotheses in both modalities, as well as the semantic and temporal compatibilities across various cross-modal hypothesis pairs during joint interpretation.
- 2) Ability to handle multiple multimodal input events in a complex input expression (e.g., a navigational inquiry that involves a composition of singular, plural, and aggregated locative references).
- 3) Avoidance from writing grammar rules as these require a high level of expertise.
- 4) Low demand on the amount training data since data collection is a costly process.
- 5) Ease of cross-modal integration as a front-end preprocess of an existing spoken dialog system, thereby enabling it to handle bimodal (speech and pen) inputs, as well as unimodal (speech-only or text-only) inputs.
- 6) Ease of portability to different information domains.

To further elaborate on the above, our approach is based on a Viterbi algorithm that enforces semantic and temporal ordering compatibilities in terms of two cost functions. Furthermore, we designed a robust interpretation framework with an integrative cost function that incorporates a weighted combination of ranked confidence scores from speech recognition, pen recognition, together with a cross-modal compatibility score. The semantic compatibility can be derived by a simple process of exploratory data analysis of a training set; hence, our approach does not involve grammar writing. The requirement on training data is relatively small, because the training data set is used only for reference and in tuning weights in the cost function. This alleviates the problem of overtraining due to insufficient training data. The cross-modality integration component identifies matched pairings between one or more spoken references with pen gestures in a multimodal expression, based on temporal and semantic compatibility. Hence, cross-modality integration can generate a unimodal (verbalized) paraphrase that is self-contained and semantically equivalent to the multimodal expression. Alternatively, cross-modality integration can also generate a verbalized paraphrase that contains meta-tags

(derived from recognized pen gestures) which represent conceptual abstractions in the place of parsed spoken locative references (e.g., the spoken reference, “*this*,” in the earlier example can be replaced with `<street:South Dongzhimen Back Street>` (i.e., 東直門南小街 in Chinese), as identified from the recognized pen stroke). Both types of paraphrase can be easily integrated with our existing spoken dialog system (SDS) [14]–[16]. The former type of paraphrase can be fed as input to the SDS but may engender redundant effort in spoken language parsing. However, this type of paraphrase eases the process of analytical comparison between unimodal and multimodal expressions, in the current investigation, e.g., analyzing relationships such as complementarity and redundancy across modalities. The latter type of paraphrase enables the carrier phrase to be interpreted by the spoken language understanding component in the SDS, while the meta-tag of `<street:South Dongzhimen Back Street>` can be directly inserted in the semantic frame as one of the key-value pairs as described in [16].¹ Thereafter, we can leverage the existing modules in the SDS for discourse inheritance, dialog modeling and response generation, which are performed based on the semantic frame. As regards portability of the proposed framework, migrating to a new information domain requires only a new domain-specific language model for the speech recognizer, as well as a set of domain-specific features list. Hence, the framework is largely domain independent. Details will be provided later.

The following presents our work in the design and collection of a multimodal corpus, characterizing speech and pen gestures for unimodal interpretation, cross-modality semantic integration, hypothesis rescoring, as well as empirical performance evaluation.

II. DESIGN AND COLLECTION OF A MULTIMODAL CORPUS

A. Information Domain

The current investigation is cast in the information domain of navigation around Beijing. Inquiries involving locative information often induce multimodal user input. We downloaded six maps from the Internet, covering five districts in Beijing. We identified about 930 locations associated with icons and labels on the maps. For each icon, we annotated their positional coordinates, corresponding to the four corners of the icon. We also categorized the icons according to “location types” and “sub-types.” There are seven location types in all, e.g., TRANSPORT, SCHOOLS_AND_LIBRARIES, etc. Each location type is further organized into 2 to 12 “subtypes.” For example, the location type TRANSPORT contains the subtypes *road*, *street*, *train_station*, *railway_station*, *railroads*, *bus_stop*, *bridge*, *intersection*, *highways*, *elevated_highway*, *elevated_road* and *road_under_construction*; while SCHOOLS_AND_LIBRARIES consists of *universities*, *institutes* and *libraries*. For a given location type and subtype, there can be multiple instances of domain-specific data entries. For example, the location type of TRANSPORT and subtype of *street* will include all the street names on the map.

¹If we follow the syntax in [16], the key-value is expressed in the form of `<street>South Dongzhimen Back Street</street>`.

TABLE I
AN ILLUSTRATIVE EXAMPLE FOR MULTIMODAL DATA COLLECTION WITH
SPEECH (\mathcal{S}) AND PEN GESTURES (\mathcal{P}). TRANSLATIONS ARE ITALICIZED

Information category: TRAVEL_TIME
Task: 告知系列你所在的位置，查詢從那裡順序到另外四所大學需要多長時間。(Specify your current location. Find the time it takes to travel to four universities of your choice.)
Multimodal input (● denotes a point and → denotes a stroke)
\mathcal{S} : 我在北郵。 "I'm at BUPT."
\mathcal{P} : ●
\mathcal{S} : 從這裡出發順序到這個大學，這個大學，這個大學，這個大學要多久？ "From here, I want to visit this university, this university, this university and this university in order. How long will it take?"
\mathcal{P} : → → → → →



Fig. 1. Data collection interface of the Pocket PC, augmented with soft buttons for logging functions (START/STOP) and loading the NEXT map. The numbers highlight some examples of location icons: 1) subject's current location (i.e., the red cross); 2) a university; 3) a road; and 4) a hospital.

We also conducted a quick survey involving ten people regarding typical inquiries from users who are trying to navigate around Beijing. These inquiries generally target nine information categories including BUS INFORMATION, TRAVEL TIME, TRANSPORTATION COSTS, ROUTE FINDING, MAP COMMANDS, etc. Based on these information categories, we designed specific tasks (31 tasks covering seven location types) such that each induces a subject to compose multimodal inquiries. Table I shows an example task and a multimodal input composed by a subject during data collection.

B. Data Collection Procedures

We invited 23 Mandarin-speaking subjects to participate in data collection. In an initial briefing session, each subject is presented with an instruction sheet listing the set of 31 tasks (as shown in Appendix B). For each task, the subject is asked to

TABLE II
EXAMPLE OF LOGGED DATA FOR MULTIMODAL INPUT BASED
ON TABLE I. EXPLANATIONS ARE IN ITALICS

Log for speech (with start and end times and the audio filename) start: 46019 end: 46030 \Program Files\DC\AudioFile10.wav
Log for pen (with each gesture numbered in order of occurrence, the recognized gesture type, start and time times and x-y coordinates of the pen down and pen up actions.)
0- point start: 46022 end: 46022 from: (152,182) to: (152,182)
1- stroke start: 46022 end: 46024 from: (152,182) to: (69,69)
2- stroke start: 46024 end: 46025 from: (69,69) to: (70,24)
3- stroke start: 46025 end: 46026 from: (69,24) to: (95,12)
4- stroke start: 46026 end: 46028 from: (93,12) to: (101,61)

formulate a multimodal input that may involve up to n locations.² The subject may refer to these locations by speech (i.e., spoken locative references) or by pen gestures. Both speech and pen inputs are recorded directly by a Pocket PC (PPC). In some of the tasks, the PPC provides contextual information "current location" with a red cross on the map. The subjects are also informed of several possible options:

- that spoken locative references may be deictic (e.g., 這裡 "here"); 這四所大學 "these four universities"); elliptic (e.g., 到這個公園要走多久 "how long does it take to walk to this park") or anaphoric (e.g., 從我的所在地到王府井要多久 "how long does it take to go from my current location to Wangfujing");
- that pen gestures may be a point, a circle or a stroke (with a pen-down gesture followed by a pen-up gesture).

Subjects are also allowed to revise and recompose their multimodal inquiries during the recording session to clearly express the intended task semantics and constraints.

C. Data Collection Setup

The recording session is carried out individually for each of the 23 subjects in an open office. The data collection setup involves a PPC with a system interface (Fig. 1). Speech input is recorded by the built-in microphone of the PPC. Pen gestures are input with a stylus. The PPC interface includes several soft buttons: The START button should be pressed to launch the automatic system logging procedure that records the speech signal, the pen gestures and the timing information between the modalities. Table II shows the logged data corresponding to the example given in Fig. 1. Pressing the NEXT button displays the map of the next task.

D. Corpus Statistics

We have collected 1518 inputs from 23 subjects in all. Among these, 1442 are multimodal and 76 are speech-only inquiries. All speech and pen data have been manually transcribed. Utterance lengths range from 2 to 54 Chinese characters, covering a vocabulary of size 521 with domain-specific named entities and spoken locative references (SLRs). A user input may consist of zero (i.e., speech only input) to six pen gestures of the types point, circle or stroke. Short inputs are typically map commands (e.g., 縮小 "zoom in"). The longest

² n is constrained to a maximum value of 6.

TABLE III
EXAMPLES FROM THE MULTIMODAL CORPUS

# Unimodal inquiries (speech only): 75	
# inquiries with spoken locative references, e.g. 我要看整個海澱區 “I wish to look at the whole <i>Haidian District</i> ”	7 (9.3%)
# inquiries without spoken locative references, e.g. 我想坐公交車 “I wish to take the bus.”	68 (90.7%)
# Multimodal inquiries: 1442	
# of inquiries with spoken references e.g. 從這個文化中心<point>到這個公園<point>要多久? “How long does it take to travel from <i>this center</i> <point> to <i>this park</i> <point>?”	1382 (95.8%)
# of inquiries without spoken references e.g. <stroke> 最快怎麼走? “<stroke> What is the fastest route?”	60 (4.2%)

input in our corpus includes several direct locative references:

我正在北京郵電大學從這裡我想依次到北京航空航
天大學中國地質大學北京科技大學和北京醫科大學可
以選擇什麼交通路線

“I’m at the *Beijing Univ. of Post and Telecommunications*.
From *here*, I want to go to the *Beihang Univ.*, *China Univ. of
Geosciences*, *Univ. of Science and Technology Beijing* and
Beijing Medical Univ. in sequence. What are the routes
available?”

We have also manually annotated the cross-modality pairings between an SLR and a pen gesture for the multimodal expressions for performance analysis. These pairings are based on human judgment (i.e., our oracle), with the objective of obtaining a holistic and coherent semantic interpretation for the bimodal input. A single SLR may map to multiple pen gestures and vice versa. For some of the SLRs or pen gestures, a mapping to the other modality cannot be found. The annotations also ignore disfluencies in the speech modality (e.g., filled pauses and repairs) and spurious gestures in pen modality (e.g., due to shaking hands). Our corpus has 3421 spoken locative references and 3590 instances of pen gestures in total. We divided the 1442 multimodal inquiries into two disjoint datasets randomly. The training set has 999 inputs and the test set has 443 inputs.

III. CHARACTERIZING SPEECH AND PEN GESTURES FOR UNIMODAL INTERPRETATION

This section describes our findings in an exploratory data analysis of the collected corpus. Our aim is to understand how individual modalities encode partial semantics that should later be conjoined to decode the holistic meaning of the user’s multimodal input. Results from the analysis are used to devise unimodal interpretation strategies for individual modalities. Table III shows some typical examples from our multimodal corpus.

A. Characterization of Spoken Inputs

The collected data offers over 3421 (count by token) and 177 (count by type) occurrences of spoken locative references (SLRs) for analysis, from which we derive the following characterizations.

- 1) **Direct references:** These involve the use of the full name of a location (e.g., 北京郵電大學 for *Beijing University of Post and Telecommunications*), its abbreviated name (e.g., 北郵 or *BUPT*), or a contextual phrase (e.g., 目前的所在地, *my current location*). Recall that the subject’s “current location” is indicated by a red cross on the map. There are 1529 occurrences of direct references involving 76 unique tokens/phrases in our corpus.
- 2) **Indirect references:** The user may also refer to a location through deixis or anaphora, e.g., 這裡 “*here*”, 那個中心 “*that center*”, 這三個商場 “*these three shopping centers*”, etc. Hence, indirect references may contain numeric features (as indicated with a numeric expression, e.g., 三 “*three*”, 幾 “*few*”, 些 “*some*”, etc.) and/or location type features (e.g., 公園 “*park*”, 大學 “*university*”). Both attributes may also be left unspecified in the SLR (e.g., 地方 “*place*”, 地點 “*location*”). The location type feature may also be ambiguous (e.g., 站 “*station/stop*”). There are 1892 occurrences of indirect references involving 101 unique SLR expressions in our corpus.

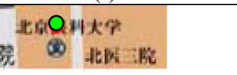





In comparison with previous work, the SLRs corresponds to the Givenness Hierarchy with four cognition statuses as mentioned in [17], where the direct references are the uniquely identifiable referents and the indirect references are the activated or familiar referents.

B. Procedure for Interpreting Spoken Locative References

Based on the above observations, we devise a three-step strategy for interpreting transcribed spoken inputs. These can be applied on manual as well as automatic transcriptions of speech.

- Step 1) **Chinese word tokenization:** The Chinese language does not have an explicit word delimiter. We perform word tokenization using a greedy algorithm with a homegrown Chinese lexicon with 43 K entries, covering nouns, verbs, phrases and SLR expressions. Should speech recognition transcripts be used, the SLR should already be tokenized based on the recognizer’s vocabulary, but may be retokenized by the current procedure.
- Step 2) **SLR Extraction:** We extract the SLR expressions by referring to our lexicon, which includes 177 unique SLR expressions. The extraction algorithm can accommodate arbitrary numeric expressions parsed from the transcribed speech. The parsed numeric expression is used to fill in the numeric feature attribute of the SLR.
- Step 3) **Hypotheses generation:** This step generates a hypothesized *list of locations* corresponding to a given SLR. A single location is typically generated for direct references, based on the name of the location or the current location from context. The list of hypothesized locations generated for an indirect reference typically includes all icons present on the map. This list may be narrowed down according to a matching location type, if the feature is specified.

TABLE IV
ILLUSTRATIONS OF THE USAGES OF DIFFERENT PEN GESTURE TYPES

Gesture	Semantics	Illustration(s)
Pointing	Indicates a single location NUM=1, e.g. a university	
Circling	A small circle indicates a single location, NUM=1, e.g. a park	
	A large circle indicates multiple locations, NUM=plural, e.g. 2 universities	
Strokes	A single stroke indicates a single location, NUM=1, e.g. a street	
	A single stroke indicates the start and end points of a path, NUM=1	
	Multiple strokes indicate a route, NUM=1, e.g. passing through 4 universities with "multiple strokes"	

Furthermore, if the numeric feature is specified, it is stored along with the generated hypothesis list. Rank ordering of the hypothesized locations is not considered for SLRs.

C. Characterization of Pen Inputs

Our corpus contains 3590 pen gestures in total. Analysis of the corpus also sheds light on the usages of the different pen gestures as illustrated in Table IV.

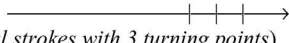
- 1) **Pointing:** This is mostly used to indicate a single location. This occurs 99.8% of the time in our corpus and the remaining occurrences are map rendering commands.
- 2) **Circling:** This includes two possible cases—small circles indicate a single location (70% of corpus statistics) and large circles indicate multiple locations (30% of corpus statistics).
- 3) **Strokes:** These include three possible cases—a stroke referring to a street or bridge (45.1% of corpus statistics), the start and end points of a path (32.3%) and multiple strokes constituting a route (22.6%).

D. Interpreting Pen Inputs

Pen inputs are interpreted based on the gesture type and its coordinates, which are compared with the positional coordinates of the icons on the map. Interpretation of each gesture type generates a ranked hypothesis list of locations, according to the following protocol.

- 1) **Point:** Icons lying within 50 pixels from the point are considered possible semantic interpretations of the gesture. These are ranked according to distances away from the point. Shorter distances give higher ranks.

Multimodal input
S: 我現在在 北郵 我要到 這四個大學 一共需要多少時間
"I am now at BUPT and I need to get to these four universities. How much time will it take?"

P: 
(multiple sequential strokes with 3 turning points)

Hypothesis lists of speech input
SLR1: **ABBREVIATION**=北郵 "BUPT"
北京郵電大學
"Beijing Univ. of Posts and Telecommunications"
SLR2: **DEICTIC**=這四個大學
"these four universities"

NUM = 4
LOC_TYPE=schools_and_public_libraries
subtype=university
中國地質大學, 北京師範大學,
北京郵電大學,

(all universities on the map shown)

Hypothesis lists of pen input (locations ranked by distance in pixels)
PenDown: TYPE=**stroke**
北京郵電大學 -1
"Beijing Univ. of Posts and Telecommunications"
西土城路 5.4 *"Xitucheng Road"*.....

TurningPt1: TYPE=**stroke**
北京航空航天大學 -1 *"Beihang Univ."*
北京航空館 5.0 *"Beijing Aviation Museum"*.....

TurningPt2: TYPE=**stroke**
中國地質大學 1.9 *"China Univ. of Geosciences"*
學院路 11.0 *"Xueyuan Road"*.....

TurningPt3: TYPE=**stroke**
北京科技大學 0.6
"Univ. of Science and Technology Beijing"
學院路 11.4 *"Xueyuan Road"*.....

PenUp: TYPE=**stroke**
北京醫科大學 -1 *"Beijing Medical Univ."*
北醫三院 7.02 *"Peking Univ. Third Hospital"*.....

Fig. 2. Illustration of the procedure for hypothesis lists generation in the speech and pen modalities, respectively.

- 2) **Circle:** The circle's area is defined by the pair of coordinates corresponding to the pen-down and pen-up gestures. Icons with overlapping areas are considered possible semantic interpretations and are ranked according to their distances away from the estimated center of the circle. Again, shorter distances give higher ranks.
- 3) **Stroke:** A hypothesis list is generated for each endpoint of a stroke, where hypotheses are ranked by distance from the endpoint. If we compare the hypothesis list of two adjacent endpoints (from one stroke or two sequential strokes) and find significant similarity (i.e., either the top three entries are identical, or the two lists have over 75% overlap), the two hypothesis lists will be merged into one according to their common entries. Using this method, we can distinguish between interpreting a *single* stroke as one location, from the other alternative of a *connecting* stroke between two locations. In the case of multiple sequential strokes, such as the three strokes in Table IV, this method enables us to interpret them as a route connecting four locations.

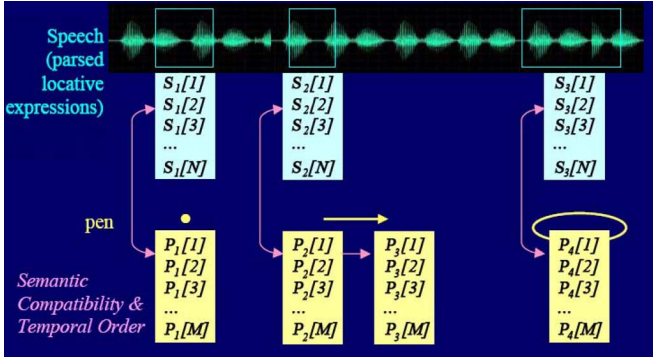


Fig. 3. Cross-modality integration procedure. Each input event (a spoken locative reference or a pen gesture such as point/circle/stroke) in each modality produces a list of hypothesized locations. There are aligned across modalities by the Viterbi algorithm while incorporating semantic compatibility and temporal order.

Fig. 2 illustrates the process of interpreting speech and pen gestures interpretation procedure is shown in Fig. 2.

IV. CROSS-MODALITY INTEGRATION

As described previously, each of the two (speech and pen) modalities abstracts the user’s intended message into a sequence of input events, i.e., in terms of spoken locative references (SLRs) or pen gestures. Each event carries semantic meaning but may contain ambiguity. The interpretation procedures for speech and pen inputs presented in (Sections III-B and III-D) derive partial semantics for each event, represented as a hypothesized list of locations. This section presents a cross-modality integration procedure that attempts to integrate the partial interpretations across modalities in order to generate a unimodal paraphrase that is semantically equivalent to the original multimodal user input. Statistics show that 67% of the multimodal inquiries in the training set have an equal number of SLR and pen gestures. However, in these cases, there may not be a one-to-one correspondence between the SLRs and pen gestures. For example:

S: 從 所在地 到 這兩個地方 要多久

P: ••

“How long will it take to travel from my current location to these two locations?”

There are two SLRs and two pointing gestures in the inquiry. However, the first SLR is an anaphora referring to the user’s current location, and the two pointing gestures both correspond to the second SLR. An overly bold assumption of one-to-one correspondence between SLRs and pen gestures can correctly interpret only 58% of the perfectly transcribed multimodal inquiries in the training set. Therefore, we perform cross-modality integration by Viterbi alignment [18] with a scoring function that enforces the temporal ordering between the sequence of SLRs and the sequence of pen gestures. The scoring function also enforces the semantic compatibility in terms of numeric (NUM) and location type (LOC_TYPE) features (see Fig. 3).

A. Enforcing Temporal Order

Analysis of our training data shows that in a multimodal input, the spoken locative reference (SLR) and pen gesture that correspond to the same intended location may not always overlap in time. In fact, the majority of cases in the training set show the pen gesture occurring either before or after its corresponding spoken reference. Hence, in the current work, we only attempt to maintain the temporal order of locative references between the speech and pen inputs. A Viterbi alignment $\mathbf{a} = a_1 a_2 a_3 \dots a_m$ can easily accommodate for this as we align the sequence of R hypothesis lists in temporal order of the SLRs $\mathbf{S} = S_1 S_2 \dots S_R$ with the sequence of Q hypothesis lists in temporal order of the pen gestures $\mathbf{P} = P_1 P_2 \dots P_Q$. Note that it is possible for a single SLR to align with multiple pen gestures (e.g., “these three universities” corresponds to three pointing inputs); as well as vice versa (e.g., “Xueyuan Road and North Huyuan Road” corresponds to a circle). The Viterbi alignment algorithm can support this by advancing the position in one hypothesis sequence (either \mathbf{S} or \mathbf{P}) while maintaining the position in the other.

B. Enforcing Semantic Compatibility

Cross-modality integration also seeks to enforce semantic compatibility. If the r th SLR is a direct reference expression, the hypothesis list S_r should contain only one element and the integration procedure seeks to match the specified location with hypotheses for the aligned pen gesture in P_q . The matching cost is defined such that if no match is found, a cost of one is incurred. If the SLR is an indirect reference expression, the hypothesis list S_r should contain multiple elements and the location type (LOC_TYPE) or numeric (NUM) features may be specified. The integration procedure checks for compatible LOC_TYPE among the hypotheses for the aligned pen gesture in P_q . A matching cost $C_M(S_r, P_q)$ of one is incurred if there is mismatch in LOC_TYPE between S_r and P_q (see (8) in the Appendix). Enforcing compatibility in NUM is a little more elaborate, especially when the value of NUM specifies multiple locations that need to be matched with the hypothesis sequences from recognized pen gestures. Hence, we use a transition cost $C_T(S_r, P_q | S_{r-i}, P_{q-j})$ which is set to the deficit in the NUM value during the transition from (S_{r-i}, P_{q-j}) to (S_r, P_q) as shown in (9) (see Appendix), where $i = \{0, 1\}$ and $j = \{0, 1\}$. The matching cost of location type and transition of numeric feature are determined with the training set. As mentioned, an SLR may align with one or more pen gestures, corresponding to one or more P_q and each may contain a different number of hypotheses. Should we encounter a tie in the conditional cumulative costs $C_C(S_r, P_q | S_{r-i}, P_{q-j})$ at (S_r, P_q) from different positions (S_{r-i}, P_{q-j}) during the course of alignment, we pick the back pointer $B(S_r, P_q)$ in the following order of precedence.

- 1) Return one step in \mathbf{P} while maintaining the position in \mathbf{S} (i.e., $i = 0$ and $j = 1$).
- 2) If the above path is not available, return one step in both \mathbf{P} and \mathbf{S} (i.e., $i = 1$ and $j = 1$).
- 3) If the above path is not available, return one step in \mathbf{S} while maintaining the position in \mathbf{P} (i.e., $i = 1$ and $j = 0$).

TABLE V
EXAMPLE ILLUSTRATING THE UNIMODAL PARAPHRASES GENERATED FROM
THE MULTIMODAL EXPRESSIONS FROM TWO DIALOG TURNS

<p>MM1: S 從 <u>我所在的地方</u> 到 <u>這裡</u> 要多久 ?</p> <p>P • (point to a hotel on the map)</p> <p>“How much time will it take to travel from <u>my current location</u> to <u>here</u>?”</p> <p>UM1: 從 <u>所在地</u> 到 <u>凱來大酒店</u> 要多久 ?</p> <p>“How much time will it take to travel from <u>my current location</u> to the <u>Gloria Hotel</u>?”</p> <p>Remarks: The system understands that “<u>我所在的地方</u>” is referring to the “current location,” which can be obtained from the dialog discourse. Also, <u>這裡</u> “here” can be jointly interpreted with the pointing gesture, due to high semantic compatibility (based on scoring). Therefore, UM1 contains the interpretations for both SLRs.</p>
<p>MM2: S 從 <u>這個酒店</u> 到 <u>這個地方</u> 有什麼車可以搭 ?</p> <p>P → (a stroke to indicate a street)</p> <p>“Which means of transportation can I use to travel from <u>this hotel</u> to <u>this place</u>?”</p> <p>UM2: 從 <u>這個酒店</u> 到 <u>王府井大街</u> 有什麼車可以搭 ?</p> <p>“Which means of transportation can I use to travel from <u>this hotel</u> to <u>Wangfujing Street</u>?”</p> <p>Remarks: The system can match <u>這個地方</u> “this place” with the stroke, due to high semantic compatibility (based on scoring). However, the indirect reference <u>這個酒店</u> “this hotel” cannot be matched with any pen gesture. Therefore, this SLR remains intact in UM2.</p>

This order aims to handle the occurrence of anaphoric reference to the user’s existing location—i.e., the anaphora does not need to pair up with a pen gesture. Details of the Viterbi algorithm are provided in Appendix A.

C. Identifying Intended Locations

This alignment procedure generates the “best” path in attempting to find an alignment between an SLR with a pen gesture in the multimodal input. The cross-modality integration procedure extracts the common location(s) found in each pair of hypothesis lists (S_r and P_q) derived from the aligned SLR and pen gesture. The number of locations extracted follows the value of the NUM feature and the ranking of locations follows those from the hypothesis list P_q from the pen modality (as described in Section III-D). The top ranking location(s) is identified as the user’s intended location(s). By substituting the identified locations in place of the SLRs in the speech input, we can generate a *unimodal, verbalized paraphrase* that is semantically equivalent to the original multimodal expression. This will be described in Section IV-E. For an *indirect* SLR that does not have any corresponding aligned pen gesture, it will remain intact in the expression and will be further disambiguated through

context inheritance in the dialog model of the SDS. An illustrative example is given in Table V.

D. Evaluating the Cross-Modality Integration Procedure

We applied the cross-modality integration procedure to both the training and test sets. Recall that thus far we have been working with hand-transcribed speech input (with perfect SLR extraction performance), together with manually annotated gesture types for pen input. The transcriptions for speech and pen are regarded as perfect. For each multimodal inquiry, we manually annotate the alignment between an SLR and a pen gesture. Based on the alignment, the user’s intended location(s) can be identified. Similarly, the Viterbi alignment is applied to each multimodal inquiry so as to obtain a system generated alignment. If the oracle and system generated alignments completely agree with each other, the multimodal inquiry is considered as correct. The cross-modality integration accuracy is defined as shown in the equation at bottom of page. The cross-modality integration procedure generated correct alignments between SLRs and pen gestures for 97.5% of training inquiries and 97.1% of the testing inquiries that contain SLR(s). The incorrect pairings shed light on possible future work, including the need to use timing information across modalities for *some* multimodal inputs; as well as the need to apply pragmatic knowledge to infer the value of the NUM feature (i.e., in the case NUM=nil) and to filter out redundant SLRs in the speech input. Further details are presented in [19].

E. Analytical Comparison Between Parallel Multimodal and Unimodal Expressions

In order to investigate the relationships between speech and pen gestures and their effects in the joint interpretation, we performed an analytical comparison between collected multimodal expressions and their automatically generated unimodal paraphrases. In order to do so, we ran the cross-modality integration procedure on the multimodal expressions. For each pair of aligned SLR and pen gesture, we can identify the user’s intended location(s). If we replace each of the SLRs with the full name of the identified location(s), we obtain the unimodal paraphrase. The correct paraphrases (over 97% of the entire data set) are extracted and combined with their semantically equivalent multimodal counterparts to form parallel corpora. More specifically, we obtain 974 multimodal and unimodal expression pairs from our training set and 430 pairs from our testing set. Comparative statistics of the multimodal and unimodal inputs are shown in Table VI. We see that the spoken components of multimodal inputs are generally shorter and cover a smaller vocabulary than their unimodal counterparts. The difference is less pronounced than expected. One reason,

Cross-modality integration accuracy

$$= \frac{\text{Total \# of multimodal inquiries with perfect match between oracle and system generated alignments}}{\text{Total \# of multimodal inquiries with SLRs}}$$

TABLE VI
PARALLEL MULTIMODAL AND UNIMODAL CORPORA STATISTICS

	Multimodal input	Unimodal paraphrase
Total # words	12,748	12,853
Average utterance length (in words)	8.8	8.9
(in chars.)	17.9	20.8
Range of utterance length (in words)	1 to 19	1 to 19
(in chars.)	2 to 54	2 to 58
Vocabulary size (# words)	473	492

TABLE VII
COMPARISONS IN PERPLEXITIES BETWEEN THE PARALLEL MULTIMODAL (MM) AND UNIMODAL (UM) INPUTS

Comparisons in Class Trigram Test Set Perplexities		
	Multimodal Input	Unimodal Paraphrases
Total # utterances	430	430
# words	4,505	4,555
Perplexity (PP)	16.5	29.5
Comparisons in Per-Utterance Perplexities		
$PP_{MM} < PP_{UM}$	356/430 utterances (82.8%)	
$PP_{MM} = PP_{UM}$	74/430 utterances (17.2%)	
$PP_{MM} > PP_{UM}$	0	

based on our observation, is the diversity of spoken deictic expressions and Chinese measure words. For example, “my current location” may be verbalized in many ways (such as 身處點, 所在地, 目前所在的地方, 現在的地方, 現在這裡, 我的位置, 我的當前位置, 當前的位置, 我現在的地方, 我現在的地點, 我當前位置, etc.).

Chinese measure words relating to location types (including 間, 個, 所, 條, 邊, 頭, 裡, 片, 帶, 塊, 點, 米, 圈, 塊兒, etc.) also contribute towards alternatives in verbalization.

We pooled the multimodal and unimodal spoken expressions together (1450 in all as presented in [19]) to train a class trigram language model. We classified the proper names (i.e., location names) into 12 equivalences classes, e.g., UNIVERSITY, HOSPITAL, STREET, etc. We also have four other equivalences classes including: ARTICLES, NUMBERS (i.e., implicit/explicit numeric expressions, e.g., 一 “one,” 幾 “few,” 些 “some,” etc.), MEASURE_WORDS and LOCATION_TYPE (e.g., the words “university,” “parks,” etc.). The language model was developed using the CMU SLM Toolkit [20]. The resulting model contains 290 unigrams, 1375 bigrams, and 2795 trigrams. The probabilities are smoothed by Katz backoff smoothing [21] with discount ratios 0.04 for unigrams, 0.36 for bigrams, and 0.38 for trigrams. The discounting thresholds for unigrams, bigrams and trigrams are 1, 5, and 7, respectively. We computed the class trigram perplexities for the multimodal and unimodal test sets, respectively. Results are shown in Table VII.

We observe that for the semantically equivalent, parallel multimodal, and unimodal corpora, the unimodal paraphrases have significantly higher perplexities. We also observe that the test set may be divided into two subsets according to comparisons in per-utterance perplexities between the multimodal (PP_{MM}) and unimodal inputs (PP_{UM}).

- 1) The subset with ($PP_{MM} = PP_{UM}$) typically contains *direct references* in speech that are semantically *redundant* with the pen modality. Each pair of (x, y) coordinates of each pen gesture in the multimodal input

TABLE VIII

EXAMPLES ILLUSTRATING PERPLEXITY REDUCTION IN DIFFERENT CASES. PERPLEXITIES OF EXAMPLES 1 AND 4 ARE THE SAME BECAUSE WE ARE USING A CLASS-BASED LANGUAGE MODEL, WHERE FULL AND ABBREVIATED NAMES OF A UNIVERSITY BELONG TO THE SAME CLASS

<p>Example 1 - $PP_{UM}=25.1$ (generated unimodal paraphrase) 從北京郵電大學到北京航空航天大學 中國地質大學 北京醫科大學 和北京科技大學要多久 “How much time will it take from <u>Beijing Univ. of Posts and Telecommunications</u> to <u>Beihang Univ.</u>, <u>China Univ. of Geosciences</u>, <u>Beijing Medical Univ.</u>, and <u>Beijing Univ. of Sci. and Technology</u>?”</p> <p>Example 2 - $PP_{MM}=5.9$ (complementarity) S: 從這裡到這四所大學要多久 P: • • • • • “how much time will it take from <u>here</u> to <u>these four univ.</u>?”</p> <p>Example 3 - $PP_{MM}=25.1$ (redundancy) S: 從北郵到北航地大北醫和北科要多久 P: • • • • • “how much time will it take from <u>BUPT</u> to <u>BUAA</u>, <u>CUG</u>, <u>BMU</u> and <u>BUST</u>?”</p> <p>Example 4 - $PP_{MM}=8.8$ (complementarity and redundancy) S: 從北郵到這四所大學要多久 P: • • • • • “how much time will it take from <u>BUPT</u> to <u>these four univ.</u>”</p>
--

matches with the direct reference to a location in the spoken utterance. Since the class-based language model gives the same probability values to the direct reference as well as the full name of the location in the unimodal paraphrase, equal per-utterance perplexities are obtained. Such redundancy is useful in real applications, where recognized transcripts may be erroneous. Redundancy across modalities motivates the use of mutual disambiguation techniques [22].

- 2) The subset with ($PP_{MM} < PP_{UM}$) typically contains *indirect references* in the speech modality that are *complementary* with the pen gestures. Either modality alone is semantically imprecise, but when their semantics are combined, the overall intended message from the user is clear. Hence, we can see that part of intended message is conveyed via the speech modality, while the remaining part is conveyed via the pen modality. The unimodal paraphrase, however, captures the full semantics of the subject’s intended message. Consequently, we obtain inequality in the perplexity values. Such complementarity offers expressive power, because the user is free to distribute various parts of the message to different modalities to ease complex communication in a succinct form, which can reduce cognitive loading for interpretation [1].

The example in Table VIII illustrates the advantage of perplexity reduction by virtue of complementarity across the speech and pen modalities, through comparison between the speech components in a multimodal expression with its counterpart in a unimodal expression. In particular, the unimodal expression in Example 1 has a perplexity of 25.1, which is reduced to 5.9 in a multimodal expression (see Example 2) with complementary speech and pen inputs. However, if the speech and pen inputs are redundant, as shown in Example 3, there is no perplexity reduction. If there is a mixture of complementary

and redundant inputs between the two modalities (see Example 4), then there is a smaller reduction in perplexity from 25.1 to 8.8. Further details of redundancy and complementarity across the speech and pen modalities, as characterized by the perplexity measure, may be found in [23].

V. HYPOTHESES RESCORING FOR ROBUSTNESS TOWARDS IMPERFECT TRANSCRIPTIONS

We attempt to extend the cross-modality integration procedure with the use of multiple recognition hypotheses in order to achieve robustness towards recognition errors. Consider the scenario in which a speech recognizer generates N -best hypotheses based on the speech input, while the pen gesture recognizer generates M -best hypotheses based on the pen input. The hypotheses are rank ordered according to their recognition scores in each individual modality. As such, we will have $N \times M$ possible candidates for cross-modality integration. In designing a rescoring mechanism for comparing these candidates for integration, we should consider such elements as the quality of the recognized spoken locative references, the quality of the interpreted pen gestures and the quality of the alignment. We will elaborate on these points in the following subsections:

A. Pruning and Scoring the Recognized Spoken Inputs

The cross-modality integration procedure has demonstrated reasonable performance in aligning spoken locative reference (SLR) expressions with pen gestures in oracle-transcribed multimodal inputs. These transcriptions are essentially perfect. However, under practical situations, captured inputs are much more problematic, due to disfluencies in the speech modality (e.g., filled pauses and repairs), spurious pen gestures and recognition errors in both modalities. These imperfections have adverse effects on cross-modality integration.

1) *Transcribing the Spoken Inputs:* We transcribed the speech signals in the multimodal corpus with a Mandarin speech recognizer [24] that is developed with the HTK toolkit [25]. This recognizer was originally trained with speech data from a general open domain. Hence, we replaced the recognizer's general-domain lexicon with a domain-specific version of 637 entries that contain names of locations in Beijing as well as frequent spoken deictic expressions. We also incorporated a domain-specific bigram language model trained from manual transcripts of the training data set. The acoustic models remain unchanged. Speech recognition performance evaluated based on the top-scoring recognition hypotheses gave overall character accuracy of 44.6%. In particular, we observe that performance is especially poor for two of the subjects who spoke Mandarin with an accent, and there was background noise. Application of the SLR extraction procedure (see Section III-A) to the top-scoring recognition hypotheses shows substitution, deletion and insertion errors in the SLRs. SLR deletion and substitution are the most prominent, frequently caused by short duration of 這兒 (meaning "here" and pronounced as /zher/) and phonetic confusion between 這 (meaning "this" and pronounced as /zhe/) and 車 (meaning "car" and pronounced as /che/). Overall, the SLR recognition accuracies (each SLR is

treated as a word)³ for the training and test sets are 38.5% and 39.3%, respectively. In other words, over half of the SLRs have not been correctly extracted. However, the majority (>60%) of the incorrectly recognized SLRs involves confusion with other SLRs carrying the same semantic meaning⁴ and hence will not affect the subsequent cross-modality integration process. Overall, 50.9% and 51.7% of the recognized SLR in training and test sets were interpreted with correct semantics.

2) *Pruning and Scoring the Spoken Inputs:* The speech recognizer may generate nonsensical hypotheses in the N -best hypothesis list. We devise a pruning strategy based on perplexity to filter out the nonsensical transcriptions. A recognition transcript with a small value of perplexity is more likely to have a reasonable interpretation. This is because the hypothesized word sequence generally conforms to the predictions by the n -gram language model. Hence, our pruning strategy targets the opposite cases—hypotheses with large perplexity values exceeding a preset threshold are filtered.

The speech component of a multimodal input expression may be transcribed by speech recognition as a hypothesized word sequence with R spoken locative references (SLRs). For a segment of the speech signal with specific start and end times, we may observe transcriptions across the N -best ($N = 100$ in this work) speech recognition hypotheses. Let S_r denotes the r th SLR in one of the speech recognition hypotheses, which is also the transcription of a specific speech signal segment. We may score the quality of this transcription by defining the normalized cost $C_S(S_r, N)$ for the recognized SLR (S_r), as shown as follows:

$$C_S(S_r, N) = 1 - \frac{n(S_r, N)}{N} \quad (1)$$

where $n(S_r, N)$ is the number of times the speech segment is transcribed as S_r across the N -best speech recognition hypotheses ($N = 100$). $n(S_r, N)/N$ is known as the N -best purity of the SLR S_r , where purity values range between 0 and 1. The higher the purity, the more preferable the SLR S_r , and the lower is the normalized cost of the speech transcription $C_S(S_r, N)$.⁵

B. Filtering and Scoring the Recognized Pen Inputs

1) *Filtering and Recognizing the Pen Inputs:* We find that subjects tend to repeat a pen gesture in referring to a location until it is highlighted on screen. We have designed a filtering mechanism to remove the repetitions. The filtering mechanism references the time and distance between two gestures. If a pen

³For each spoken input expression, we compare the list of parsed SLR(s) from its oracle transcription with the list of parsed SLR(s) from its speech recognition transcription. The SLR recognition accuracy is defined as



$$\text{SLR Recognition Accuracy} = \frac{N_{\text{SLR}} - I_{\text{SLR}} - S_{\text{SLR}} - D_{\text{SLR}}}{N_{\text{SLR}}}$$

where N_{SLR} is the total number of SLRs in the oracle transcriptions; I_{SLR} , S_{SLR} and D_{SLR} are the numbers of insertion, substitution, and deletion errors from the speech recognition transcriptions, respectively.

⁴The confusion between SLRs during speech recognition may involve only the measure word and hence does not alter the semantic meaning.

⁵It is conceivable that should a pen gesture recognizer be used to generate M -best recognition hypotheses, a similar M -best purity may be incorporated in the cost function for the pen modality.

TABLE IX
ILLUSTRATIVE EXAMPLES ON THE RECOGNITION
ERRORS OF CIRCLE AND STROKE

A flat circle is mis-recognized as “stroke”.	
A distorted stroke with high ROC, which is rejected by the recognizer.	

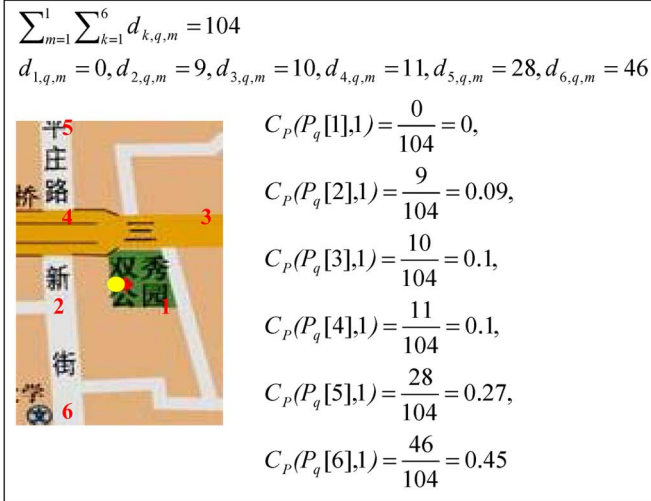


Fig. 4. Illustrative example for the calculation of normalized cost for an interpreted pen gesture.

gesture shows the x and y coordinates within a short amount of time and a short distance, the later one is filtered out. We have developed a pen gesture recognizer, based on a simple algorithm that proceeds through a sequential procedure of recognizing a point, a circle and a stroke, as follows.

- 1) **Recognizing Points:** If the pixel distance between the pen down and pen up coordinates is fewer than q ($= 6$ pixels), the input is considered as a point. Detected pointing actions with temporal difference less than 0.25 s are considered repetitive and the redundancy will be discarded. If the pen gesture is not classified as a point, it will be evaluated as a circle or stroke, described as follows.
- 2) **Recognizing Circles:** A pen gesture is recognized as a circle if 80% of its x and y coordinates appears at least twice and if it contains no more than two convex hulls. If the pen gesture is not classified as circle, it will be evaluated as a stroke, described as follows.
- 3) **Recognizing Strokes:** Since strokes are directional, a pen gesture is recognized as stroke if one or both of the x and y coordinates shows directional migration towards pen down coordinates, also the radius of curvature cannot exceed a preset threshold [5]. If the pen gesture is not classified as stroke, it will be rejected.

This simple pen gesture recognition algorithm can only generate a single output hypothesis. Further extension can be made to generate M -best pen gesture hypotheses. Overall pen gesture recognition accuracy is 86.6%. Table IX shows some pen gesture recognition errors. Among the incorrectly

recognized pen gestures, they contain confusions may carry the same semantic meaning and here the pen recognition error will not affect the subsequent integration process. Overall, 91.3% of the recognized pen gestures can be interpreted with correct semantic meaning.

2) *Rescoring the Pen Inputs:* A multimodal input expression may be transcribed as a sequence of Q pen gestures with recognized pen gesture type. Each is interpreted as a list of hypothesized locations, i.e., P_q for the q th pen gesture in the input expression. The interpretations are based on locations on the map that lie within a maximum distance d_{max} (empirically set at 50 pixels based on training data) from the coordinates of the pen gesture and are rank ordered based on these distances $d_{k,q,m}$, where k indexes the hypothesized locations in P_q and may range from 1 to K_q ,⁶ and m indexes the number of recognized pen gesture types and $M = 1$ in current work. To score a particular interpretation $P_q[j]$ in the hypothesized list, we define the normalized cost of interpretation for the pen modality $C_P(P_q[j], M)$ as shown in (2). The smaller the distance $d_{j,q}$, the lower the normalized cost $C_P(P_q[j], M)$ and the more preferable the interpretation for the pen gesture. The normalized costs of the K_q hypothesized locations in P_q will sum to 1. Fig. 4 shows an illustrative example of the normalized costs of different interpretations of a pointing gesture. Hypothesized locations for the circle must have their coordinates enclosed by the circle. The locations are rank ordered based on their distances away from the circle's center.

$$C_P(P_q[j], M) = \frac{d_{j,q,m}}{\sum_{m=1}^M \sum_{k=1}^{K_q} d_{k,q,m}} \quad (2)$$

C. Pruning and Scoring Cross-Modality Integrations

The cross-modality integration procedure described in Section IV incorporates a simple cost function for the Viterbi algorithm that penalizes for mismatches in directly referenced locations, LOC_TYPE and NUM features. High accuracies in cross-modality alignment were obtained based on near-perfect multimodal input transcriptions. However, in handling the imperfect N -best speech recognition and M -best pen recognition outputs, we need to enforce tighter constraints on semantic compatibility. We have established by the perplexity measure (in Section IV-E) that direct references should be *semantically redundant* with the corresponding pen gestures. Additionally, indirect references should be *semantically compatible* with their corresponding pen gestures. Hence, we propose to incorporate a *pruning mechanism* for candidate integrations which involve mismatches in locations between interpreted pen gestures and direct references in speech, or mismatches in the LOC_TYPE and NUM features between interpreted pen gestures and indirect references in speech. Table X presents an illustrative example. The top-scoring speech recognition hypothesis contains the direct reference 北醫 (BMU, Beijing Medical University) while the second best contains 北郵 (BUPT, Beijing University of Post and Telecommunications) instead. However, since the corresponding pen gesture (first gesture) is a point with positional

⁶ K_q is empirically set at 10 for all q , based on analysis of the training data.

TABLE X
ILLUSTRATIVE EXAMPLE OF THE PRUNING MECHANISM FOR
CANDIDATES FOR CROSS-MODALITY INTEGRATIONS

First best speech recognition hypothesis (<i>pruned because of the mismatch in location between the first interpreted pen gesture and the first direct SLR</i>)	
s: 出北醫 要到地大 到北科大 到北航 最後到 哪兒 北醫 坐什麼車	
p:	• • • • •
Second best speech recognition hypothesis	
s: 出北郵 到地大 到北科大 到北航 最後到 哪兒 北醫 坐什麼車	
p:	• • • • •
Interpretation of the first pen gesture	
Point	
北京郵電大學(BUPT)	$d_1=0$
西土城路(West Tucheng Road)	$d_2=18$
學院南路(Xueyuan South Road)	$d_3=27.79$
北京師範大學(Beijing Normal University)	$d_4=31$

coordinates that coincide with the BUPT icon (such that the distance $d_1 = 0$), the cross-modality integration between the top-scoring speech recognition hypothesis and the pen gesture is pruned.

Candidate integrations that survive the pruning mechanism will each have a Viterbi alignment cost $C_A(S_R, P_Q)$, which is computed with alignment and transition costs described in Section IV based on the pair of hypothesis lists (S_R, P_Q) as defined in (9) in Appendix A. S_R is the hypothesized transcription of the speech input that contains R recognized spoken locative references. P_Q is the hypothesized transcription of the pen input that contains Q interpreted pen gestures. We define the normalized cost of integration $C_I(S_R, P_Q)$, where the subscript I denotes “integration,” as shown in (3). $\max\{C_A\}$ is the maximum possible Viterbi alignment cost that is empirically obtained from training data

$$C_I(S_R, P_Q) = \frac{C_A(S_R, P_Q)}{\max\{C_A\}},$$

where

$$0 \leq C_I(S_R, P_Q) \leq 1. \quad (3)$$

D. Rescoring Cross-Modality Integrations

Recall that in the current work, the speech recognizer is set to generate N -best hypotheses ($N = 100$) and the pen gesture recognizer generates only the top-scoring gesture type ($M = 1$). Cross-modality integration begins with a pruning process (see Section V-C). Surviving candidates (pairs of recognized speech and recognized pen hypothesis) are rescored with the following procedures.

- 1) For each candidate, we apply cross-modality integration to its pair of hypothesis lists (S_R, P_Q) . Should these include incompatible semantics, the candidate is pruned. If the candidate survives, we compute its normalized cost of integration $C_I(S_R, P_Q)$ based on (3).
- 2) We focus on the hypothesized transcription of the pen input P_Q . For each of the Q interpreted pen gestures (indexed by q), we select the interpretation j_q that is semantically compatible with its aligned SLR and compute the normalized cost of pen interpretation $C_P(P_q[j_q])$ [see (2)]. Should

there be multiple semantically compatible interpretations, their normalized costs are summed. The overall cost of interpreted pen gestures for P_Q is defined as

$$C_P(P_Q) = \frac{1}{Q} \sum_{q=1}^Q C_P(P_q[j_q], M). \quad (4)$$

- 3) We focus on the hypothesized transcription of the speech input S_R . For each of the R recognized SLRs (indexed by r), we compute its normalized cost of recognized SLR, i.e., $C_S(S_r, N)$ [see (1)], which is derived from the N -best purity. The overall cost of recognized SLR for S_R is defined as

$$C_S(S_R) = \frac{1}{R} \sum_{r=1}^R C_S(S_r, N). \quad (5)$$

- 4) The rescoring function that is used to evaluate each candidate for cross-modality integration is a linear combination of the three normalized cost functions relating to the alignment, interpreted pen gestures, and recognized SLRs, i.e.,

$$C_{Tot}(S_R, P_Q) = w_I C_I(S_R, P_Q) + w_P C_P(P_Q) + w_S C_S(S_R),$$

where

$$0 < w_I, w_P, w_S < 1$$

and

$$w_I + w_P + w_S = 1. \quad (6)$$

We select values for the weights w_I , w_P , and w_S , by grid search to maximize cross-modality alignment accuracies based on the training data. The values selected are $w_I = 0.5$, $w_P = 0.35$, and $w_S = 0.15$. The “optimized” weight of the pen modality is higher than that of speech modality, possibly due to higher pen gesture recognition accuracies, as compared with the speech recognition accuracies. All candidates for cross-modality integration are rescored according to (6) and reranked in ascending order of scores. The candidate with minimum overall cost $C_{Tot}(S_R, P_Q)$ is identified as the preferred cross-modality alignment.

E. Evaluating the Rescoring Procedure

The application of the rescoring procedure to the candidate hypotheses for cross-modality integration has brought some improvements to the alignment accuracies in the training and test sets of our multimodal corpus. Table XI summarizes the results of the percentage of correctly aligned expressions. These are expressions for which our framework can generate unimodal verbalized paraphrases that are semantically equivalent with the original multimodal expressions. Improvements in integration accuracies brought about by cross-modality hypotheses rescoring is statistically significant from 54.8% to 69.9% in test set results ($\alpha = 0.01$, one-tailed z -test). Further analysis of our results (see Table XII) shows that there can be correct cross-modality integration despite recognition errors in speech and/or pen modalities. The N -best hypothesis rescoring framework can effectively rerank the hypothesis pairs to obtain

TABLE XI

PERFORMANCE OF THE CROSS-MODALITY INTEGRATION, MEASURED IN TERMS OF % OF CORRECTLY ALIGNED EXPRESSIONS IN THE TRAINING AND TEST SETS

	Training Set	Test Set
# expressions	957	425
Cross-modality integration of oracle transcriptions in both modalities based on <u>temporal order only</u> (i.e. align one-by-one)	58.0%	58.3%
Cross-modality integration of oracle transcriptions in both modalities based on the Viterbi Alignment in Section IV.	97.1%	97.5%
Cross-modality integration of top-scoring speech and pen input recognition hypothesis based on <u>temporal order only</u>	28.9%	27.3%
Cross-modality integration of top-scoring speech recognition hypothesis and recognized pen inputs based on the Viterbi Alignment in Section IV and in the Appendix A	53.7%	54.8%
Top candidate obtained after cross-modality integration and rescoring of the N -best speech recognition outputs ($N=100$) with the first best recognized pen inputs hypotheses.	67.5%	69.9%

TABLE XII

DETAILED PERFORMANCE STATISTICS OF THE *test set*

Pen recognition	SLR recognition	#inquiries in the test set (425 in total)	Correct integration with top-scoring hypotheses from each modality	Correct integration with N -best ($N=100$) speech recognition hypotheses and M -best ($M=1$) pen recognition hypotheses
Correct	Correct	96/425 (22.6%)	96/96 (100%)	96/96 (100%)
Correct	Incorrect	256/425 (60.2%)	92/256 (35.9%)	152/256 (59.4%) ⁷
Incorrect	Correct	40/425 (9.4%)	31/40 (77.5%)	32/40 (80%) ⁸
Incorrect	Incorrect	33/425 (7.8%)	14/33 (42.4%)	17/33 (51.5%) ⁹
Overall			54.8%	69.9%

correct integration, as illustrated by the examples in Table XIII.

In addition, analysis of the incorrect alignments (after rescoring and reranking) suggests that the incorporation of finer cross-modality timing information will be helpful. Such timing information should be used judiciously since the modalities are not necessary simultaneous and user's integration pattern may vary during the interaction [12]. Furthermore, a good number of the errors are associated with the SLR 這裡 "here" having an unspecified NUM feature and can thus be aligned with an



⁷Improvements in integration accuracies brought about by cross-modality hypotheses rescoring is statistically significant from 35.9% to 59.4% in the presence of speech recognition errors ($\alpha = 0.01$, one-tailed z -test).

⁸Improvements in integration accuracies brought about by cross-modality hypotheses rescoring is statistically insignificant in the presence of pen recognition errors ($\alpha = 0.05$, one-tailed z -test).

⁹Improvements in integration accuracies brought about by cross-modality hypotheses rescoring is statistically significant from 42.4% to 51.5% in the present of both speech and pen recognition errors ($\alpha = 0.05$, one-tailed z -test).

TABLE XIII

EXAMPLES ON THE CORRECT INTEGRATION WITH THE PRESENT OF SLR AND/OR PEN RECOGNITION ERROR

<p>Example 1 (with SLR recognition errors): Reference transcriptions: S:從 這兒 到 這四個大學 要多久? "How much time will it take from <u>here</u> to <u>these four universities</u>?" P: • • • • •</p>	
<p>Top-scoring speech and pen recognition hypotheses: S:從 這裡 到 這些地方 要多久? "how much time will it take from <u>here</u> to <u>these locations</u>?" P: • • • • •</p>	
<p>Remark: the reference SLR, 這裡, has the same semantic meaning as 這兒 (i.e. "here") and does not affect the subsequent cross modality integration. The numeric and the location type features are lost during recognition of the second SLR "these locations". The proposed framework can find out the correct alignment and extract the name of the four universities based on the complementary relation between the modalities.</p>	
<p>Example 2 (with pen gesture recognition error): Reference transcriptions: S:在 這裡 逛一圈要多久? "how long will it take to stroll around <u>here</u>?" P: ○ (user drew a flat circle to indicate a street)</p>	
<p>Top-scoring speech and pen recognition hypotheses: S:在 這裡 逛一圈要多久? P: → (pen gesture mis-recognized as a stroke to indicate a street)</p>	
<p>Remark: The pen interpretation method in Section III.4 can identify the street as indicated by the mis-recognized pen gesture and hence the recognition error does not affect the cross-modality semantic integration process.</p>	
<p>Example 3 (with SLR and pen gesture recognition errors): Reference transcriptions: S:這個公園 什麼時候 開放 "what is the opening hours of <u>this park</u>?" P: • (a big point within the icon of a park)</p>	
<p>Top-scoring speech and pen recognition hypotheses: S:這兒 公園 什麼時候 開放 "what is the opening hours of <u>here</u> park?" P: ○ (a circle within the icon of a park)</p>	
<p>Remark: Although the numeric and location type features are missed in the recognized SLR and the point is mistaken as circle by the pen gesture recognizer, the framework can integrate the two modalities correctly and identify the park indicated by the user.</p>	

arbitrary number of pen gestures. Making the assumption of NUM=1 should be helpful for error recovery.

Analysis of the incorrect interpretations found that deficiency in the timing information, handling of the unspecified numeric feature (e.g., 這裡 "here" has NUM=nil and can be aligned with any number of pen gesture instances without penalty in the alignment cost) are the two main causes. Incorporation of the timing information can help to reduce the association between the SLR and pen gesture with temporal difference $\geq \theta$. Generation of specific numeric feature can provide a more specific alignment cost.

VI. CONCLUSION AND FUTURE WORK

We present a framework pertaining to automatic semantic interpretation of multimodal user interactions using speech and

pen gestures. The two input modalities (speech and pen) abstract the user's intended message differently into input events, i.e., key terms/phrases in speech and different gestures in the pen modality. The semantics of an input event may be imprecise, incomplete, or erroneous due to misrecognitions. The proposed framework begins by generating (partial) interpretations for each input event, which are represented as a ranked list of hypothesized interpretations. We devise a *cross-modality semantic integration procedure* to align input events in the speech modality with those in the pen modality using the Viterbi algorithm. Cost functions are designed to enforce the constraints of temporal ordering of the input events in each modality, as well as the semantic compatibility between hypothesized interpretations across modalities. Hence, the alignment integrates across modalities and disambiguates among possible interpretation alternatives to decode the user's holistic communicative intent. We designed and collected a multimodal corpus in domain of city navigation to support our investigation. This corpus contains many multimodal expressions with frequent locative references. The speech and pen modalities have been transcribed by hand. The overall speech character recognition and pen gesture type recognition accuracies are 44.6% and 89.9%, respectively. Application of cross-modality integration to these near-perfect transcripts generated correct unimodal paraphrases for over 97% of the training and testing sets. However, if we replace with the top-scoring speech and pen recognition transcripts, the performance drops to 53.7% and 54.8% for the training and test sets, respectively. In order to achieve robustness towards imperfect transcripts, we extend our framework with a *hypothesis rescoring procedure*. For each multimodal expression, this procedure considers all candidates for cross-modality integration based on the N -best ($N = 100$) speech recognition hypotheses and the M -best ($M = 1$) pen input recognition hypotheses. Note that the single recognized pen gesture can generate Q location hypotheses that are fed into the cross-modality hypothesis rescoring procedure [see (2)]. Rescoring combines such elements as the integration scores obtained from the Viterbi algorithm, N -best purity for recognized spoken locative references, as well as distances between coordinates of recognized pen gestures and relevant icons on the map. Experiments using the N -best ($N = 100$) speech recognition hypothesis and top-scoring ($M = 1$) pen recognition hypotheses show that the rescoring and reranking helped improve the performance of correct cross-modality interpretation to 67.5% and 69.9% for the training and testing sets, respectively. We expect that further performance gains will be achieved if we incorporate the use of a speech recognition lattice in this work,¹⁰ as well as extend our pen recognizer to produce multiple hypotheses. Correct cross-modality semantic integration enables our framework to the multimodal input expression to be paraphrased as a unimodal (speech-only) input, for subsequent processing of our existing dialog system with natural language generation and dialog and discourse modeling components. Hence, the cross-modality semantic integration framework offers an elegant front-end extension to our dialog system, to enable it to handle both uni-

modal (speech-only) as well as multimodal (speech and pen) inputs. Future work includes the investigation of cross-modality timing information to aid semantic interpretation, the handling of ellipsis in the speech component of a multimodal expression and identifying possible cross-modality correlation patterns that may help improve performance in multimodal semantic interpretation.

APPENDIX A VITERBI ALIGNMENT ALGORITHM

Notations

S_r	List of hypothesis of the r th SLR.
P_q	List of hypothesis of the q th pen gesture instance.
$C_M(S_r, P_q)$	Matching cost between S_r and P_q .
$C_T(S_r, P_q S_{r-i}, P_{q-j})$	Transition cost from (S_{r-i}, P_{q-j}) to the current position (S_r, P_q) . It indicates the deficit in the NUM value for $i, j = \{0, 1\}$.
$C_A(S_r, P_q)$	Cumulative cost (the best partial alignment) up to the position of (S_r, P_q) from (S_1, P_1) .
$C_C(S_r, P_q S_{r-i}, P_{q-j})$	is the conditional cumulative cost at (S_r, P_q) from the position (S_{r-i}, P_{q-j}) for $i, j = \{0, 1\}$, such that $C_C(S_r, P_q S_{r-i}, P_{q-j}) = C_M(S_r, P_q) + C_A(S_r, P_q) + C_T(S_r, P_q S_{r-i}, P_{q-j})$.
$B(S_r, P_q)$	Back pointer of the position (S_r, P_q) determined by the local minimization of $C_A(S_r, P_q)$.
$\Psi(r, q)$	Backtracking path obtained from the back pointer $B(S_r, P_q)$.
$C_A(S_r, P_q)$	Cumulative cost at the final position (S_r, P_q) .
R	Total number of SLRs in the inquiry.
Q	Total number of pen gesture instances in the inquiry.

Initialization

$$C_A(S_1, P_1) = C_M(S_1, P_1)$$

$$B(S_1, P_1) = nil.$$

$$\text{Recursion } (\forall (r, q) = \{(1, 1), \dots, (R-1, Q), (R, Q-1), (R-1, Q-1)\})$$

$$C_M(S_r, P_q) = \begin{cases} 0 & S_r \cap P_q \neq \emptyset \\ 1 & S_r \cap P_q = \emptyset \end{cases} \quad (7)$$

$$C_T(S_r, P_q | S_{r-i}, P_{q-j}) = \text{deficit in the NUM value for } \{i, j\} = \{(0, 1), (1, 0), (1, 1)\} \quad (8)$$

¹⁰Additional experiments on the test set show that when $N = 10$, correct cross-modality integration is 58.3% that compares with 69.9% when $N = 100$.

$$C_A(S_r, P_q) = \begin{cases} C_C(S_r, P_q | S_r, P_{q-1}), & r = 1 \\ C_C(S_r, P_q | S_{r-1}, P_q), & q = 1 \\ \min\{C_C(S_r, P_q | S_{r-1}, P_{q-1}), \\ C_C(S_{r-1}, P_q), C_C(S_r, P_q | S_r, P_{q-1})\}, & \text{otherwise} \end{cases}$$

$$B(S_r, P_q) = \arg \min_X \{C_A(S_r, P_q)\}$$

for $X = \{(r-1, q), (r, q-1), (r-1, q-1)\}$.

Termination:

$$C_A(S_R, P_Q) = \min \{C_C(S_R, P_Q | S_{R-1}, P_{Q-1}), C_C(S_R, P_Q | S_{R-1}, P_Q), C_C(S_R, P_Q | S_R, P_{Q-1})\}$$

$$B(S_R, P_Q) = \arg \min_X \{C_A(S_R, P_Q)\}$$

for $X = \{(R-1, Q), (R, Q-1), (R-1, Q-1)\}$.

Path Backtracking:

while $i > 0$, do $\{\Psi(r[j], q[j]) := B(S_{r[i]}, P_{q[i]}), i := j\}$
 for $(r[j], q[j]) = \{(R, Q), \dots, (1, 1)\}$.

APPENDIX B

EXAMPLES OF TASKS, LOCATION TYPES, AND SUBTYPES OBTAINED FROM DATA COLLECTION

TABLE B.1
 INFORMATION CATEGORY: ROUTE_FINDING

Information category: ROUTE_FINDING	
Task: Inquire about the route to walk through the <u>Palace Museum</u> , the <u>Great Hall of the People</u> and the <u>Military Museum of Chinese People's Revolution</u> .	
SLRs: the <u>Palace Museum</u>	
LOC_TYPE: PUBLIC_FACILITIES_AND_SERVICES subtype: <i>heritage</i>	
LOC_TYPE: LEISURE_FACILITIES subtype: <i>museum</i>	
SLR: the <u>Great Hall of the People</u>	
LOC_TYPE: POLITICAL_FEATURES subtype: <i>district_office</i>	
LOC_TYPE: LEISURE_FACILITIES subtype: <i>theater</i>	
SLR: the <u>Military Museum of Chinese People's Revolution</u>	
LOC_TYPE: LEISURE_FACILITIES subtype: <i>museum</i>	
Remark: A location may classify into multiple location type and subtype; a location type and subtype include multiple locations.	

TABLE B.2
 INFORMATION CATEGORY: TRAVEL_TIME

Information category: TRAVEL_TIME
Task: Tell the system that you are now at the <u>Beijing University of Posts and Telecommunications</u> . Inquire about the route from the <u>Beijing University of Posts and Telecommunications</u> to the <u>Beihang University</u> , the <u>China University of Geosciences</u> , the <u>University of Science and Technology Beijing</u> , the <u>Beijing Medical University</u> in order
SLRs: the <u>Beijing University of Posts and Telecommunications</u> , the <u>Beihang University</u> , the <u>China University of Geosciences</u> , the <u>University of Science and Technology Beijing</u> and the <u>Beijing Medical University</u>
LOC_TYPE: SCHOOL_AND_PUBLIC_LIBRARIES subtype: <i>university</i>
Remark: A task may contain up to 6 locations.

The full set of tasks, information categories, location types, and subtypes can be found in our project website: <http://www.se.cuhk.edu.hk/~pyhui/multimodal.htm>.

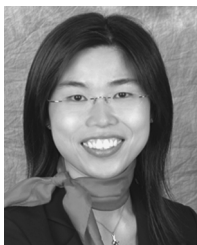
ACKNOWLEDGMENT

The authors would like to thank Dr. Frank Soong and the late Dr. Jianlai Zhou for their guidance for the first author in Mandarin speech recognition during her summer internship at MSRA. This work is affiliated with the CUHK MoE-Microsoft Key Laboratory of Human-Centric Computing and Interface Technologies. In addition, we would like to thank the anonymous reviewers for their helpful comments.

REFERENCES

- [1] S. Oviatt, R. Coulston, and R. Lunsford, "When do we interact multimodally? Cognitive load and multimodal communication patterns," in *Proc. ICMI*, 2004, pp. 129–136.
- [2] S. Oviatt *et al.*, "Designing the user interface for multimodal speech and pen-based gesture applications: State-of-the-art systems and future research directions," *Human-Comput. Interaction*, pp. 263–322, 2000.
- [3] A. G. Hauptmann, "Speech and gestures for graphic image manipulation," in *Proc. CHI*, 1989, pp. 241–245.
- [4] L. Nigay and J. Coutaz, "A generic platform for addressing the multimodal challenge," in *Proc. CHI*, 1995, pp. 98–105.
- [5] S. Wang, "A multimodal galaxy-based geographic system," S.M. thesis, Massachusetts Inst. Technol., Cambridge, 2003.
- [6] M. Johnston *et al.*, "Unification-based multimodal integration," in *Proc. COLING-ACL*, 1997, pp. 281–288.
- [7] M. Johnston, "Unification-based multimodal parsing," in *Proc. COLING-ACL*, 1998, pp. 624–630.
- [8] L. Wu *et al.*, "Multimodal integration—A statistical view," *IEEE Trans. Multimedia*, vol. 1, no. 4, pp. 334–341, Dec. 1999.
- [9] W. Wahlster *et al.*, "SmartKom," Germany [Online]. Available: www.smartkom.org.
- [10] M. Johnston and S. Bangalore, "Finite-state multimodal parsing and understanding," in *Proc. COLING*, 2000, pp. 369–375.
- [11] J. Chai *et al.*, "A probabilistic approach to reference resolution in multimodal user interfaces," in *Proc. IUI*, 2004, pp. 70–77.
- [12] J. Chai *et al.*, "Optimization in multimodal interpretation," in *Proc. ACL*, 2004, pp. 1–8.
- [13] S. Qu and J. Chai, "Salience modeling based on non-verbal modalities for spoken language understanding," in *Proc. ICMI*, 2006, pp. 193–200.
- [14] S. F. Chan and H. Meng, "Interdependencies among dialog acts, task goals and discourse inheritance in mixed-initiative dialog," in *Proc. HLT*, 2002, pp. 197–202.
- [15] Z. Y. Wu, H. Meng, H. Ning, and C. F. Tse, "A corpus-based approach for cooperative response generation in a dialog system," in *Proc. ISCSLP*, 2006, pp. 614–626.

- [16] H. Meng and D. Li, "Multilingual spoken dialog systems," in *Multilingual Speech Processing*, T. Schultz and K. Kirchhoff, Eds. New York: Academic, 2006, pp. 399–447.
- [17] A. Kehler, "Cognition status and form of reference in multimodal human–computer interaction," in *Proc. AAAI*, 2000, pp. 685–690.
- [18] P. Brown *et al.*, "The mathematics of statistical machine translation: Parameter estimation," *Comput. Linguist.*, vol. 19, no. 2, pp. 263–311, 1993.
- [19] P. Y. Hui and H. Meng, "Joint interpretation of input speech and pen gestures for multimodal human–computer interaction," in *Proc. Interspeech*, 2006, pp. 1197–1200.
- [20] P. Clarkson and R. Rosenfeld, "Statistical language modeling using the CMU-Cambridge toolkit," in *Proc. Eurospeech*, 1997, pp. 2707–2710.
- [21] J. T. Goodman, "A bit of progress in language modeling: Extended version," Microsoft Research, Tech. Rep. MSR-TR-2001-72, 2001.
- [22] S. Oviatt, "Multimodal system processing in mobile environments," in *Proc. UIST*, New York, 2000, pp. 21–30.
- [23] P. Y. Hui *et al.*, "Complementarity and redundancy in multimodal user inputs with speech and pen gestures," in *Proc. Interspeech*, 2007, pp. 2205–2208.
- [24] E. Chang *et al.*, "Large vocabulary mandarin speech recognition with different approaches in modeling tones," in *Proc. ICSLP*, 2000, pp. 983–986.
- [25] "HTK Speech Recognition Toolkit." Univ. of Cambridge, U.K. [Online]. Available: <http://htk.eng.cam.ac.uk/>



Pui-Yu Hui (GS'08) received the B.S. and M.Phil. degrees in systems engineering and engineering management from The Chinese University of Hong Kong (CUHK) in 2000 and 2003, respectively. She is currently pursuing the Ph.D. degree in the same department.

She is an Assistant Computer Officer with the CUHK MoE-Microsoft Key Laboratory of Human-Centric Computing and Interface Technologies. Her research interests include spoken document retrieval and multimodal input integration

and understanding.



Helen M. Meng (M'99) received the S.B., S.M., and Ph.D. degrees, all in electrical engineering, from the Massachusetts Institute of Technology.

She has been a Research Scientist with the MIT Spoken Language Systems Group, where she worked on multilingual conversational systems. She joined The Chinese University of Hong Kong (CUHK) in 1998, where she is currently Professor in the Department of Systems Engineering and Engineering Management and Associate Dean of Research of the Faculty of Engineering. In 1999, she established

the Human–Computer Communications Laboratory at CUHK and serves as Director. In 2005, she established the Microsoft-CUHK Joint Laboratory for Human-Centric Computing and Interface Technologies, which was upgraded to MoE Key Laboratory in 2008, and serves as Co-Director. Her research interest is in the area of human–computer interaction via multimodal and multilingual spoken language systems, as well as translanguing speech retrieval technologies.

Helen serves as the Editor-in-Chief of the IEEE TRANSACTIONS ON SPEECH, AUDIO, AND LANGUAGE PROCESSING. She is also a member of Sigma Xi and the International Speech Communication Association.