Latent Semantic Analysis for Multimodal User Input With Speech and Gestures

Pui-Yu Hui and Helen Meng

***Abstract*—This paper describes our work in semantic interpretation of a "multimodal language" with speech and gestures using latent semantic analysis (LSA). Our aim is to infer the domain-specific informational goal of multimodal inputs. The informational goal is characterized by lexical terms used in the spoken modality, partial semantics of gestures in the pen modality, as well as term co-occurrence patterns across modalities, leading to "multimodal terms." We designed and collected a multimodal corpus of navigational inquiries. We also obtained perfect (i.e. manual) and imperfect (i.e. automatic via recognition) transcriptions for these. We automatically align parsed spoken locative references (SLRs) with their corresponding pen gesture(s) using the Viterbi alignment, according to their numeric and location type features. Then, we characterize each cross-modal integration pattern as a 3-tuple multimodal term with SLR, pen gesture type and their temporal relationship. We propose to use latent semantic analysis (LSA) to derive the latent semantics from manual (i.e. perfect) and automatic (i.e. imperfect) transcriptions of the collected multimodal inputs. In order to achieve this, both multimodal and lexical terms are used to compose an inquiry-term matrix, which is then factorized using singular value decomposition (SVD) to derive the latent semantics automatically. Informational goal inference based on the latent semantics shows that the informational goal inference accuracy of a disjoint test set is 99% and 84% when a perfect and imperfect projection model is used respectively, which performs significantly better than (at least 9.9% absolute) the baseline performance using vector-space model (VSM).**

***Index Terms*—Multimodal user interfaces, speech recognition, gesture recognition, latent semantic analysis.**

I. Introduction

MULTIMODAL user interfaces aim to mimic the human's multi-sensory perceptual capabilities to be able to process combined input modalities, including speech, gestures, touch, gaze, etc. In particular, much attention is devoted to spoken language combined with gestures. Spoken language offers a rich modality for expression and so does the gesture modality (e.g. in pointing, drawing, writing, highlighting, soft keyboard typing, etc.). Speech is a fast medium for verbalized communication and provides a hands-free interface, but may be affected by ambient noise. Gestures offer convenience in conveying spatial or graphical information, preserve privacy and are unaffected by noise. Consider the multimodal input: "What street is this? $<$ PEN STROKE $>$," the message components in individual modalities are semantically inexact. However, the multimodal expression captures precise semantics clearly and concisely in its entirety. Human interpretation of multimodal expressions can readily combine the analyses from both modalities to decode the intended message. Furthermore, combined cross-modal analysis can achieve robust interpretation in face of signaling errors or signal degradations due to adverse environmental conditions (e.g. overly bright/dark illumination, ambient noise, jerky motions, etc.) Studies have also shown that multimodal communication can minimize human cognitive load in communication. Additionally, the growing penetration of mobile information appliances (especially smartphones) and advancements in wireless connectivity are driving the proliferation of ubiquitous, media-rich information services. This creates a dire need for multimodal speech and gestural interfaces. All these advantages provide strong impetus for research in computational frameworks for semantic interpretation of multimodal expressions.

Automatic semantic interpretation of multimodal input is an active research area. Previous approaches largely involve partial interpretations of each modality and their integration to decode the user's message. However, such approaches are hampered by several challenging issues: First, there is considerable variability in how the user distributes the message across modalities. Contributing factors include the nature of the message and task at hand, the interface, environmental conditions and the user's (idiosyncratic) preferences. Second, input events that have cross-modal semantic correlations may not have temporal synchrony. Contributing factors are similar to those mentioned above. Third, message components in individual modalities may be irregular, imprecise or incomplete. Speech may contain disfluencies such as false starts, repairs, truncations and filled or unfilled pauses. There may also be hyper-articulated or slurred productions, prolonged/reduced segments and ungrammatical structures. Gestures may be imprecise, jittery, repetitive or spurious (i.e. unrelated to the message). Successive gestures may even be contiguous. Fourth, speech and pen recognition technologies cannot guarantee perfect performance. They are particularly error-prone in face of adverse environments and signal irregularities. Fifth, processing signals from multiple modalities inevitably increases the number of dimensions in computing. These factors present high demands for empirical data that is needed to model cross-modal characteristics.

Manuscript received March 06, 2013; revised July 29, 2013; accepted November 12, 2013. Date of publication December 11, 2013; date of current version January 08, 2014. This work supported in part by a grant from the HK SAR Government Research Grants Council (Project No. CUHK415609). The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Gary Geunbae Lee.

The authors are with the Human-Computer Communications Laboratory, The Chinese University of Hong Kong, Shatin NT, Hong Kong (e-mail: pyhui@se.cuhk.edu.hk; hmmeng@se.cuhk.edu.hk).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TASLP.2013.2294586

In order to address the above research issues, we propose to adopt the perspective of bimodal speech and gestural input as a "multimodal language," and investigate the use of latent semantic analysis (LSA) that can *jointly* identify cross-modal integration patterns in the form of "key multimodal terms" and decode their message semantics for inferring the user's informational goal. The cross-modal integration patterns unite correlated spoken reference(s) and gesture(s) in a multimodal expression. The problem is especially complicated for expressions with multiple references. This is because a spoken reference may map to zero, one or more gestures, and vice versa. The five issues mentioned above further confound the problem. This work presents a data-driven framework for extracting cross-modal integration patterns that incorporate their combinatoric and temporal relations. We have devised an approach that applies the Viterbi algorithm to align spoken references and gestures pointing to various map locations. The alignment enforces semantic constraints based on numeric expressions parsed from speech, spatial vicinity obtained from the map, as well as temporal ordering of the spoken and gestural events. This method presents relatively low requirements on the amount of semantic labeling. The alignment is used to generate *multimodal terms* that capture the user's cross-modal integration patterns. Semantic decoding should utilize these patterns in the computation of a holistic interpretation of the entire multimodal expression. We propose to adopt LSA to capture and represent the contextual semantics of key multimodal terms for interpretation of multimodal inputs, relating to inference of the user's informational goal. LSA presents a principled, representational model for contextual semantics of verbal entities in a corpus. It has been successfully applied to many problems in language processing and text mining. Analysis involves projection onto a low-dimensional latent space that describes contextual semantics. We will leverage this parsimonious description of salient cross-modal patterns in providing constraints for decoding multimodal expressions. The constraints should also facilitate robust decoding through cross-modal reinforcement, disambiguation and error compensation.

This paper is organized as follows: Section 2 is a brief literature review of multimodal speech and gestural interfaces and LSA. Section 3 describes the information domain in which we conduct our multimodal investigation. Sections 4 and 5 present the multimodal corpus that we have designed, collected and annotated to support our investigation, as well as an analysis of typical cross-modal groupings that capture salient structural semantic relations pertaining to concepts and informational goals in the information domain. Section 6 sketches the latent semantic model for multimodal semantic interpretation, targeting the inference of the user's informational goal. Section 7 presents the implementation of the vector-space model for informational goal inference, to serve as the reference baseline in performance evaluation. Sections 8 and 9 respectively discuss the use of perfect and imperfect transcriptions (for speech and pen inputs) in semantic inference with latent projections. Section 10 presents an analysis of the latent semantic space used in informational goal inference. Finally, Section 11 presents the conclusions and future directions.

II. Previous Work

A. Multimodal Speech and Gestural (SG) Interfaces

Multimodal SG interfaces have been an active research area. A pioneering effort is the Quickset system [1], [2] that runs on a handheld PC for military and medical informatics. There are also previous work such as MIT's multimodal Galaxy-based geographical system [3]; AT&T's MATCH kiosk multimodal city guide [4]; SmartKom Mobile for ubiquitous information access [5]; Microsoft's MiPAD for personal information assistance [6]; the RealHunter system for real-estate information [7]; a three-dimensional decorating domain [8]; as well as the combined use of handwriting and speech for high performance dictation [9], [10]. The W3C Multimodal Interaction Working Group has also developed the EMMA (Extensible Multimodal Annotation Markup Language) that specifies annotations for multimodal user input, in order to extend the Web infrastructure with speech and pen/mouse/keystroke event-driven capabilities.

Fusion techniques for multimodal SG interfaces primarily operate at the semantic level. The main approaches include:

(i) Frame-based heuristic integration–An attribute-value data structure is used to represent partial semantics from each input modality and each type of contextual information. The data structures are then merged according to top-level control heuristics and pattern matching techniques. For example, [11] devised the "melting pot" representation with a three-step procedure handling simultaneous input, sequential input and context-based fusion. [3] developed a multimodal context resolution module for resolving anaphoric and deictic references [12] based on syntax and semantics in spoken language.

(ii) Typed feature structures with unification-based parsing [13]–The N-best speech/gesture recognition hypotheses are represented as typed feature structures (FS). Temporally compatible multimodal combinations are combined by multi-dimensional chart parsing using a declarative unification-based grammar.

(iii) Combined confidence scoring–Generalized posterior probabilities are computed for recognition hypotheses in one modality as constrained by contextual information from the other [9]; or a weighted combination of likelihood scores from each recognizer is used for multimodal recognition [10].

(iv) Hybrid symbolic-statistical approach [14]–This approach filters for semantically plausible associations between modalities and posterior probabilities from different recognizers are integrated through a hierarchy to form a combined recognition decision. Another implementation in SmartKom [15] applies a unification approach on the recognition hypotheses graphs for each modality, followed by adaptive confidence rescoring.

(v) Multimodal weighted finite-state transducers (WFSTs) [16], [17], [18]–WFSTs offer tight coupling across modalities such that a gestural input can dynamically alter the language model for speech recognition.

(vi) Probabilistic graph-matching [7], [19]–Attribute relational graphs are used to represent multimodal input and context information. Each graph node encodes semantic / temporal information and each edge encodes semantic / temporal relations. The probabilistic graph matching algorithm is an optimization algorithm that attempts to find the best match among graphs to satisfy all relevant constraints.

(vii) Statistical classification-based approach [20], [21]–recognition outputs based on speech and head motions that relate to "agree" versus "disagree" are fed into a Support Vector Machine (SVM) to perform binary classification [20]. Fusion between gesture and speech utilizes a gesture-based salience-driven language model in speech recognition in [21]. The work develops further into classification of user intentions. Multiple SVMs are built for all pairs of classes to combine binary for multi-class user intention categorization. Instance-based classification based on the similarity between a testing instance and closest training instances (k-nearest neighbors) is also used with Hamming distances between nominal semantic features and phoneme features.

Semantic decoding of multimodal input is an exciting area with many open problems. A critical factor in cross-modal fusion that needs further research is a descriptive framework for cross-modal integration patterns that incorporate their combinatoric and temporal relations. This work extends previous research in handling multimodal, speech and pen-based user input expressions, which are navigational inquiries, composed of singular, plural or aggregated locative references. To derive the mutual correspondences between an arbitrary number of spoken and gestural locative references in an input expression, we need to decode various one-to-many or many-to-one semantic mappings between them. We will describe a novel cross-modal integration approach that applies the Viterbi algorithm to align spoken references and gestures pointing to various map locations. We then apply LSA as a representational model to capture the contextual semantics in a reduced dimensional space for robust interpretation of multimodal inputs, relating to inference of the user's informational goal.

B. Latent Semantic Analysis (LSA)

LSA is a mathematical technique that can automatically extract the relations between contextual usage of words and related documents. A comprehensive account of its principles and applications may be found in [22], [23]. LSA originated from the field of information retrieval [24], [25], [26], [27], where a major concern is to match the words in a query with words in a document collection in order to retrieve relevant documents. However, the problem is confounded by features such as polysemy and synonymy in words (or lexical terms). Some approaches towards information retrieval hence retarget the underlying topical meaning in queries and documents and seek to match at the abstract concept level. The LSA technique assumes the presence of underlying latent semantics, which are derived from correlation patterns between words and documents. Computation for LSA begins by forming a word document matrix with each row representing a unique word and each column representing a unique document. For a given document collection, this matrix is conceivably large and sparse. LSA then applies singular value decomposition (SVD) to decompose the large matrix into the product of three matrices. This decomposition enables the projection of the original word-document space into a latent space of significantly reduced dimensions. The respective vector representations of words and documents can also be projected into this latent space as new vectors of reduced dimensionality. This latent space offers a parsimonious description of the salient semantic associations among words and documents, which can be automatically derived from corpus statistics. The advantages offered by LSA in generating a representational model for human verbal concepts has been applied to a host of problems [23]. Examples include word clustering [28], topic clustering, language modeling in speech recognition [29], [30], [31], information filtering [32], [33], information routing [34], [35], spoken interface control [36] and other applications. In this project, we propose to investigate the use of latent semantic modeling for multimodal concepts, in order to enable semantic decoding of multimodal expressions.

III. Information Domain

This work focuses on the city navigation domain for Beijing.[1] We conducted a quick survey involving 10 people regarding typical inquiries from users who are trying to navigate around a city with Chinese speech and mobile device with a map displayed on screen. These inquiries generally target 9 informational goals, including: BUS_INFORMATION, CHOICE_OF_VEHICLE, MAP_COMMANDS, OPENING_HOURS, RAILWAY_INFORMATION, ROUTE_FINDING, TIME_CONSTRAINT, TRANSPORTATION_COSTS and TRAVEL_TIME. We designed 31 specific tasks (leading to 66 inquiries) based on these informational goals where the tasks will induce the subject to compose speech-based multimodal inquiries during data collection. The inquiries are allowed to cover up to six locations. Table I shows an example of a task, which requests a subject to compose a multimodal inquiry to specify his current location and ask about the travel time to four other universities of his choice.

IV. Design and Collection of a Multimodal Corpus

We invited 23 Mandarin-speaking subjects to participate in data collection. They are presented with an instruction sheet listing the set of tasks. The subject is asked to formulate a multimodal inquiry for each task. The subject may refer to these locations, i.e. spoken locative references (SLRs), by Mandarin speech or pen gestures. SLRs may be deictic (e.g. 這裡 *"here"*, 這四所大學 *"these four universities"*), elliptic (e.g. 到這個旅館要走多久？ *"how long does it take to walk to this hotel?"*) or anaphoric (e.g. 從所在地到故宮要多久？ *"how long does it take to go from my current location to the Palace Museum?"*). Pen gestures may be a point, a circle or a stroke. Both speech and pen gestures are recorded directly by a Windows CE mobile device. In some of the tasks, the Windows CE mobile device provides contextual information of the current location with a red cross on the map. Subjects are allowed to revise and recompose their inquiries during the recording to clearly express the intended task semantics.

[1] Beijing is selected due to the ready availability of comprehensive maps online at the time of this investigation.

TABLE I
EXAMPLE OF A MULTIMODAL INPUT WITH SPEECH ($\boldsymbol{S}$) AND PEN GESTURES ($\boldsymbol{P}$). THE PEN GESTURES INCLUDE A POINT (IN RED) AND A LONG STROKE (IN GREEN), WHICH ARE ILLUSTRATED IN FIG. 1. TRANSLATIONS ARE PROVIDED IN ITALICS

INFORMATIONAL GOAL: TRAVEL_TIME
TASK: 告知系統你所在的位置，查詢從那裡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: $\boldsymbol{S}$: 我在 這個地方 從 這裡 出發 順序 到 這四個大學 要多久？ $\boldsymbol{P}$: • ———————— (a point) (a long stroke connecting 5 locations) "*I'm at this place. From here, I want to visit these four universities in order. How long will it take?*"

TABLE II
EXAMPLES OF COLLECTED MULTIMODAL INQUIRIES SHOWING THAT A SINGLE SLR CAN MAP TO MULTIPLE PEN GESTURES (TOP) AND VICE VERSA (BOTTOM). TRANSLATIONS ARE PROVIDED IN ITALICS

A SINGLE SLR MAPPING TO MULTIPLE PEN GESTURES:
$\boldsymbol{S}$: 我在 這個地方 順序 到 這四個大學 要多久？ $\boldsymbol{P}$: • •••• (a point) (four points) *"I'm at this place. I wish to visit these four universities in order. How long will it take?"*
A SINGLE PEN GESTURE MAPPING TO MULTIPLE SLRS:
$\boldsymbol{S}$: 請問新東安、東方及賽特的營運時間？ $\boldsymbol{P}$: ○ (a circle covering three shopping centers) *"May I know the opening hours of the Xindong`an Plaza, the Oriental Plaza and the Scitech Plaza?"*

The recording session is carried out individually in an open office (i.e. with background noise). Speech input is recorded by the built-in microphone of a Windows CE mobile device (e.g. Pocket PCs, Windows phones, etc. and we use a Pocket PC in this work). Pen gestures are input with a stylus. The subject needs to press the start and stop buttons on the user interface to launch and stop the automatic logging procedure respectively. Specifically, the stop button appears on screen after the start button has been pressed. The procedure records the speech, pen gestures and their mutual timing information. The subject can then press the next button to display the map of the next task. Each subject is asked to formulate 66 multimodal inquiries related to *all* the 31 tasks.

We have collected 1,518 inquiries from 23 subjects. Among these, 1,442 are multimodal and 76 are monomodal (speech-only) inquiries. All speech and pen gestures have been manually transcribed by a research staff member. Utterance lengths range from 2 to 54 Chinese characters, covering a vocabulary of size 521 with domain-specific named entities and SLRs. The shortest and longest inquiries collected are, respectively, map commands (e.g. 縮小, or "*zoom in"*) and multimodal inquiries with the full name of several universities. A user input may consist of zero (i.e. speech-only inquiry) to six pen gestures, which may be a point, a circle or a stroke. We have also manually annotated the cross-modal alignment between an SLR and a pen gesture for the multimodal inquiries. The annotation will serve as the gold standard in experimentation. The alignments are based on human judgment, with the objective of obtaining a holistic and coherent semantic interpretation for the bimodal inquiry. Specifically, during a recording session, manual transcription is done simultaneously. The subject is then requested immediately after recording to label the alignments based on the manual transcripts. The transcriptions ignore disfluencies in the speech modality (e.g. repairs) and spurious gestures in the pen modality (e.g. jittery hands). The collected corpus contains 3,421 SLRs and 3,590 instances of pen gestures in total. We randomly divide the 1,442 multimodal inquiries into disjoint training and test sets in a 7:3 ratio. The training and test sets have 999 and 443 inquiries respectively.

We have also transcribed the speech signals using the multilingual Google Speech Input API.[2] This speech recognition engine primarily performs Mandarin recognition and generates output in Simplified Chinese (which is the setting we selected). However, we notice that the engine may also generate English words in its recognition output. The recognizer can generate n-best recognition results ($n \leq 20$) for an input spoken utterance. Speech recognition performance evaluated based on the top-scoring recognition hypotheses gave overall character accuracy of around 78%.[3] We have also 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 [37]. This simple pen gesture recognition algorithm can generate m-best output hypotheses. Overall pen gesture recognition accuracy is 86.6% for top-ranking hypotheses.

V. CORPUS ANALYSIS

A single SLR may map to multiple pen gestures and vice versa (see Table II). Some of the SLRs or pen gestures may not find a mapping to the other modality (e.g. ellipsis[4] and anaphora[5]). The following describes our findings in corpus analysis. Results from the analysis are used to devise the automatic cross-modal alignment strategies.

[2] Google Speech Input API http://code.google.com/chrome/extensions/experimental.speechInput.html

[3]

$$C = \frac{N_{char} - I_{char} - S_{char} - D_{char}}{N_{char}}$$

where C denotes the character recognition accuracy, N_{char} is the total number of characters in the manual transcriptions; I_{char}, S_{char} and D_{char} are the numbers of insertion, substitution and deletion errors from the speech recognition transcriptions respectively.

[4] An example of *ellipsis* is: "→ 最快路線" (meaning "*the fastest route*"), when a single pen stroke connects multiple locations. The subject wishes to find the fastest route connecting the indicated locations. Ellipsis occurs here because the spoken utterance omits mentioning the locations.

[5] Anaphora refers to "the use of a pronoun or similar word instead of repeating a word used earlier" [38] where the interpretation of an anaphora can be from the same input, contextual information or dialog history. The anaphora is underlined in: "從 所在地 到 這兩個地方•• 要多久" *("How long will it take to travel from my current location to these two locations?")*, and the other SLR corresponds to the two pointing gestures.

A. Spoken Inputs

Analysis of spoken inputs shows that subjects may refer to a location directly or indirectly. (1) Direct references involve the use of full name of a location (e.g. 故宮博物院, or "*the Palace Museum*"), its abbreviated name/alias (e.g. 故宮, 紫禁城, or "*the Forbidden City*") or a contextual phrase (e.g. 所在地, or "*current location*," which is indicated by a red cross on the map). There are 1,529 occurrences of direct references involving 76 unique tokens. (2) Indirect references involve the use of deixis or anaphora, e.g. 這裡 (meaning "*here*") and 這三個商場 (meaning "*these three shopping plazas*"). Hence, indirect references may contain numeric (e.g. 三, or "*three*," and 些, or "*some*") and/or location type features (e.g. 公園, or "*park*," and 大學, or "*university*"). Both features may be left unspecified (e.g. 地方, or "*place*") or ambiguous (e.g. 站, or "*station/stop*"). There are 1,892 occurrences of indirect references involving 101 unique tokens in our corpus.

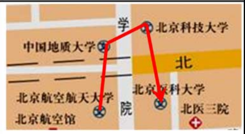
We parse the spoken inputs for SLRs based on the above observations using a greedy algorithm, which can accommodate arbitrary numeric expressions. The parsed numeric and location type expressions are used to fill in the numeric and location type feature attributes of the SLR. Afterwards, we interpret the SLR according to its reference type: a single location is generated from a direct reference, while a list of locations is generated from an indirect reference–which includes all icons (with matching location type) present on the map.

For each input spoken utterance, we parse the SLRs from the oracle (i.e. manual) transcription and the speech recognition transcripts separately, in order to assess the effect of speech recognition errors on SLR extraction. Based on the training set, we observe that speech recognition errors led to errors in one-third of the SLRs. The majority of these are SLR deletion errors (94%) and the remaining are substitution errors. Among the SLR deletion errors, over half of them are caused by heavy co-articulation of 這兒, which is an SLR meaning "*here*." These two Chinese characters respectively have the pronunciations /*zhe*/ and /*er*/, but when they occur together, they tend to be heavily co-articulated to become /*zher*/. The duration of this production is short (typically around 0.15 second) and weak. The recognizer misrecognizes it as noise for half of their occurrences. Stronger productions of /*zher*/ have also been observed to be misrecognized as the English word "*chart*." Similarly, some productions of 這裡 /*zhe li*/ (also meaning "*here*") are misrecognized as the English word "*cherry*." In these cases, the misrecognition leads to SLR deletion. SLR substitution errors are mainly caused by the misrecognition of Chinese measure words (e.g., 這三所大學 is recognized as 這三個大學, meaning "*these three universities*"). There are also two SLR insertion errors, caused by the misrecognition of 請問 (pronounced as /*qing wen*/, meaning "*I would like to know*") as 崇文 (pronounced as /*chong wen*/, which is a name of a district in Beijing).

B. Pen Gesture Inputs

The training set of our corpus contains 2,502 pen gesture instances, which include 1,863 pointing, 460 circling and 179 stroke gestures. Analysis of the corpus also sheds light on the usages of the different pen gestures as illustrated in Table III. The pointing gesture (i.e. POINT) is mostly used to indicate a single location. This occurs 99.8% (1859/1863) of the time in our training corpus. The circling gesture (i.e. CIRCLE) includes two possible cases: a small circle indicating a single location (70%, 322/460) and a large circle indicating multiple locations (30%, 138/460). Strokes (i.e. STROKE) include three possible cases: a stroke referring to a single location (45.3%, 81/179), the start and end points of a path (32.4%, 58/179) and a long stroke constituting a route (22.3%, 40/179).

TABLE III
ILLUSTRATIONS OF THE USAGE OF DIFFERENT TYPES OF PEN GESTURES

Gesture	Semantics	Illustration(s)
POINT	Indicates a single location, NUM=1, e.g. a university	
CIRCLE	A small circle indicates a single location, NUM=1, e.g. a park	
	A large circle indicates multiple locations, NUM=plural, e.g. two universities	北京科技大学 中国地质大学 北
STROKE	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 at each end point	
	A long stroke with one or more turning points indicates a route, NUM=1, e.g. a long stroke passing through four universities	 北京科技大学 中国地质大学 北 北京航空航天大学 北京航空馆 北医三院

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. Shorter distances are given higher ranks. For the pointing gesture, the locations are ranked according to distances away from the point. For the circling gesture, the locations are ranked according to distances away from the estimated center of the circle. A hypothesis list is generated for each endpoint/turning point of a stroke.

C. Temporal Relationships

Temporal relationships between a pair of corresponding SLR and pen gesture(s) include simultaneous occurrences (i.e. with temporal overlap, denoted as "SIM") and sequential occurrences (i.e. without temporal overlap, denoted as "SEQ"). Further analysis shows that most of the subjects have adopted predominantly (with over 70% consistency) either simultaneous (87%, 20/23) or sequential (8.7%, 2/23) temporal patterns between speech and pen gestures [39], [40].

D. Cross-modal Alignment

We automatically align SLR(s) with pen gesture(s) based on two constraints: (1) temporal ordering and (2) semantic com-

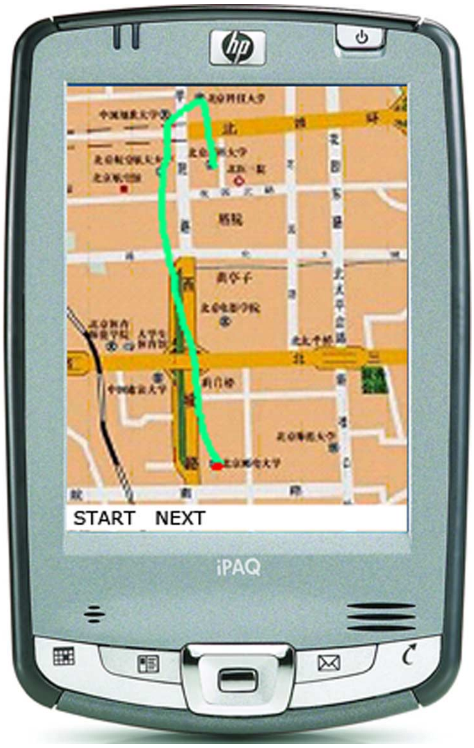


Fig. 1. An illustration of the multimodal input shown in Table I. The dot is shown in red while the stroke is highlighted in green.

patibility, by means of a Viterbi alignment algorithm [41]. We enforce temporal ordering since SLR and pen gesture(s) are not guaranteed to overlap in time. Therefore, we align the sequence of R hypothesis lists in temporal order of the SLRs $S = S_1\ S_2 \ldots S_R$ with the sequence of Q hypothesis lists in temporal order of the pen gestures $P = P_1\ P_2 \ldots P_Q$. In order to enforce semantic compatibility between SLR(s) and pen gesture(s), our approach checks the location type and numeric features of an SLR. If the rth SLR is a direct reference, the hypothesis list S_r should contain only one element and the integration procedure seeks to match the specified location with hypotheses for the qth pen gesture in the alignment. A cost is incurred if no match is found. If the SLR is an indirect reference expression, the hypothesis list S_r should contain one or more elements and the location type or numeric features may be specified. The integration procedure checks for the compatible location type feature among the hypotheses for the aligned pen gesture in P_q. A cost of one is incurred if there is mismatch in location type feature between S_r and P_q. Compatibility in numeric feature is ensured by checking whether an SLR is associated with a compatible number of pen gesture(s). For cases with a one-to-many mapping between an SLR and its corresponding pen gesture(s), the pen gestures are indicated as a group (i.e. MULTI-POINT, MULTI-CIRCLE and MULTI-STROKE) during cross-modal alignment. We use a transition cost which is set to the deficit in the numeric value during the transition from a pair of SLR and pen gesture(s) to another. This is used to indicate that there are too few or too many pen gestures aligned with one SLR or vice versa. The detailed explanation of this process of Viterbi alignment is provided in [37]. An illustrative example is shown in Fig. 2.

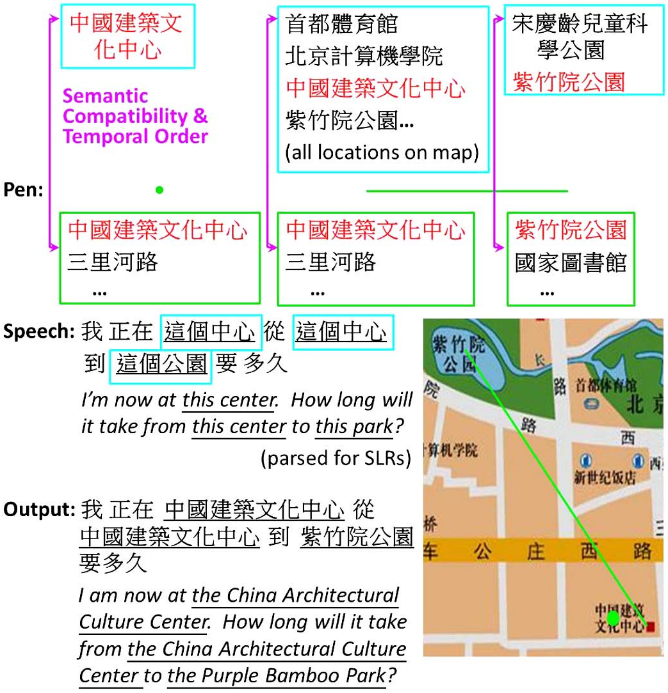

中國建築文化中心
Semantic Compatibility & Temporal Order
首都體育館
北京計算機學院
中國建築文化中心
紫竹院公園...
(all locations on map)
宋慶齡兒童科學公園
紫竹院公園
Pen:
中國建築文化中心
三里河路
...
中國建築文化中心
三里河路
...
紫竹院公園
國家圖書館
...
Speech: 我 正在 這個中心 從 這個中心 到 這個公園 要 多久
I'm now at this center. How long will it take from this center to this park?
(parsed for SLRs)
Output: 我 正在 中國建築文化中心 從 中國建築文化中心 到 紫竹院公園 要多久
I am now at the China Architectural Culture Center. How long will it take from the China Architectural Culture Center to the Purple Bamboo Park?
紫竹院公园
首都体育馆
北京
计算机学院
新世纪饭店
车 公 庄 西 路
中国建筑文化中心

Fig. 2. An example illustrates Viterbi alignment between a spoken Mandarin utterance and pen gestures that consist of a point and a stroke. We first parse the spoken input for referential and locative expressions. Then a hypothesized list of locations for each SLR (S_r) is generated. At the same time, we compute the distance between the coordinates of the pen gestures (P_q) and icons on the map to generate a ranked list of locations. Thereafter, Viterbi alignment that incorporates constraints of semantic compatibility and temporal ordering integrates the input from both modalities. Translations are provided in italics.

TABLE IV
STATISTICS OF UNIQUE LEXICAL AND MULTIMODAL TERMS (COUNTED BY TYPE)

# of Multimodal terms	567
# of Lexical terms	314
Total number of terms	881

E. Cross-modal Integration Patterns

The Viterbi alignment algorithm gives us a grouping of an SLR, its aligned pen gesture and their temporal relationship (either simultaneous sim or sequential seq, as mentioned in Section V.3). We present the grouping as a multimodal term, which is a 3-tuple that consists of corresponding SLR(s) and pen gesture(s), together with their temporal relationship. Multimodal terms generated from the example in Fig. 2 include: <這個中心|POINT|SIM> “$<$ *this center*|POINT|SIM $>$,” <這個中心|STROKE|SIM> “$<$ *this center*|STROKE|SIM $>$ ” and <這個公園|STROKE|SIM> “$<$ *this park*|STROKE|SIM $>$.”

A lexical term refers to a tokenized Chinese word from the spoken input which is not an SLR. Lexical terms tokenized from the example in Fig. 2 include: 我 (“*I*”), 正在 (“*currently at*”), 從 (“*from*”), 到 (“*to*”), 要 (“*require*”) and 多久 (“*how long*”). Statistics of the lexical and multimodal terms (counted by type) are shown in Table IV. Among the 2,429 multimodal terms found in the training set (with both SLR and pen gesture), 74% (1797/2429) are simultaneous and 26% (632/2429) are sequential.

For multimodal terms (e.g. ellipsis and anaphora) for which a mapping to the other modality cannot be found, the cross-modal temporal relationship indicated by "$\emptyset$." For example, consider the multimodal input with elliptic locative references (i.e. the SLR is omitted in speech):

$\boldsymbol{S}$: 開放時間 *Opening hours.*

$\boldsymbol{P}$: •••

Here, the user is asking for the opening hours of three places. The corresponding multimodal term is: $<\emptyset|$MULTI-POINT$|\emptyset>$.

We have applied the cross-modality alignment procedure mentioned in Section V.4 to multimodal inquiries of both training and test sets. These are transcribed manually and automatically (with speech and pen recognition). The inquiry-based cross-modality integration accuracy is defined in terms of the fraction of the number of inquiries in the dataset with perfect match between the oracle- and system-generated alignments.

Comparison between the system-generated correspondences with the manually-annotated correspondences shows that the cross-modality alignment procedure can generate correct alignments between SLRs and pen gestures for 97.2% (971/999) of training inquiries and 97.5% (432/443) of the inquiries in test set with speech and pen inputs against perfect, manual transcriptions. This corresponds respectively to 98.3% (2387/2429) and 98.6% (971/985) of multimodal terms with correct alignments.

Analogous numbers for speech and pen inputs against imperfect, automatic transcriptions are 68.3% (682/999) correctly aligned inquiries in the training set and 61.9% (274/443) in the test set. This corresponds respectively to 74.4% (1807/2429) and 67.3% (663/985) multimodal terms with correct alignments (see Table V). Serious deletion errors of SLRs in the automatic transcriptions have adverse effect on cross-modal alignment.

TABLE V
PERFORMANCE OF THE CROSS-MODALITY ALIGNMENT MEASURED IN TERMS OF THE PERCENTAGE OF CORRECTLY ALIGNED INQUIRIES AND MULTIMODAL TERMS IN THE TRAINING AND TEST SETS

	Training Set	Test Set
Number of inquiries	999	443
Number of multimodal terms	2429	985
Number of manually transcribed inquiries with correct correspondence between SLRs and pen gestures	97.2% (971/999)	97.5% (432/443)
Number of automatically transcribed inquiries (i.e. recognized speech and pen inputs) with correct correspondence between SLRs and pen gestures	68.3% (682/999)	61.9% (274/443)
Number of terms in manual transcriptions with correct correspondence between SLRs and pen gestures	98.3% (2387/2429)	98.6% (971/985)
Number of terms in automatic transcriptions (i.e. recognized speech and pen inputs) with correct correspondence between SLRs and pen gestures	74.4% (1807/2429)	67.3% (663/985)

VI. LATENT SEMANTIC MODEL

We apply LSA on multimodal inputs for multimodal semantic interpretation. LSA can capture regularities in lexical and multimodal terms, in relation to their usage contexts. A key usage context is the informational goal of the user's input expression.

Based on the training set, we summarize the associations between M terms (including both lexical and multimodal terms) and N inquiries in a term-inquiry matrix G of dimensions $M \times N$. Each column represents an inquiry and each row represents a term. The element $g_{m,n}$, is the weight of the term m in the nth inquiry. We obtain $g_{m,n}$ based on the term frequency normalized for inquiry length and term entropy [42] for both multimodal and lexical terms, as shown in Equation (1). The concept behind $g_{m,n}$ is similar to TF-IDF [43]:

$$G = \begin{bmatrix} g_{1,1} & \cdots & g_{1,n} & \cdots & g_{1,N} \\ \vdots & \ddots & \vdots & \iddots & \vdots \\ g_{m,1} & \cdots & g_{m,n} & \cdots & g_{m,N} \\ \vdots & \iddots & \vdots & \ddots & \vdots \\ g_{M,1} & \cdots & g_{M,n} & \cdots & g_{M,N} \end{bmatrix} \tag{1}$$

where $g_{m,n} = (1 - \varepsilon_m)\dfrac{\kappa_{m,n}}{\lambda_n}$,

$$\varepsilon_m = -\frac{1}{\log N}\sum_{n=1}^{N}\frac{\kappa_{m,n}}{\tau_m}\log\frac{\kappa_{m,n}}{\tau_m}, \tag{2}$$

$\kappa_{m,n}$ denotes the number of times the term m occurs in the nth inquiry, λ_n is the total number of terms in the nth inquiry, ε_m denotes the normalized entropy of term m in the training set; and τ_m is the total number of times that term m occurs in the training set.

We apply SVD to the term-inquiry matrix G and decompose it into a product of three matrices of order R, as shown in Equation (3). Associations between terms and latent semantics are summarized in matrix U, and the associations between inquiries and latent semantics are represented in matrix V. Each column of U contains the estimated weight of each term m that corresponds to the latent semantic category r, while each column of V^T contains the estimated weight of each inquiry n that corresponds to the latent semantic category r. Equation (3) projects the space of terms and inquiries onto an R-dimensional space which is defined by the orthonormal basis given by the column vectors u_m and v_n from matrices U and V respectively. We can then derive the projection from the term-inquiry matrix G to the latent semantic-inquiry matrix W from Equation (3) and perform the projection using Equation (4).

$$\begin{aligned} G &= USV^T \\ &= \begin{bmatrix} u_{1,1} & \ldots & u_{1,R} \\ \vdots & \ddots & \vdots \\ u_{M,1} & \ldots & u_{M,R} \end{bmatrix} \begin{bmatrix} s_{1,1} & 0 & 0 \\ 0 & \ddots & 0 \\ 0 & 0 & s_{R,R} \end{bmatrix} \begin{bmatrix} v_{1,1} & \ldots & v_{1,R} \\ \vdots & \ddots & \vdots \\ v_{N,1} & \ldots & v_{N,R} \end{bmatrix}^T \\ &= UW \end{aligned} \tag{3}$$

$$\Rightarrow U^T G = U^T U W = W \tag{4}$$

where U is the left unitary matrix of dimensions $M \times R$, S is the diagonal matrix of singular values sorted in descending order of dimensions $R \times R$, V^T is the right unitary matrix of dimensions $R \times N$, $R = \min\{M, N\}$ is the order of decomposition, W is latent semantic-inquiry matrix of dimensions $R \times N$; and superscript T is the transpose of the matrix.

Recall that we need to project the latent semantic-inquiry matrix W to the space of informational goals. Hence, we need to find the association between informational goals and latent semantics, which can be done using Equation (5).

$$FW = H \tag{5}$$

where F is goal-latent semantic matrix of dimensions $A \times R$, and H is the goal-inquiry matrix of dimensions $A \times N$.

The goal-inquiry matrix H can be obtained from the training set by summarizing the associations between the A informational goals and N inquiries. Each column represents an inquiry and each row represents the weights corresponding to an informational goal. This is presented in Equation (6):

$$H = [\, h_1 \quad \ldots \quad h_N \,] = \begin{bmatrix} h_{1,1} & \ldots & h_{1,N} \\ \vdots & \ddots & \vdots \\ h_{A,1} & \ldots & h_{A,N} \end{bmatrix} \tag{6}$$

where A is the total number of informational goals within the application; and h_n is the vector of weights for each of A informational goals corresponding to the nth inquiry.

Based on the manually labeled training set, h_n is a binary vector whose element corresponding to the labeled informational goal is 1 and the remaining elements are 0.

We can then obtain the goal-latent semantic matrix F using Equation (7), which is derived from Equation (5).

$$\begin{aligned} &FW = H \\ &\quad \Rightarrow FWW^T = HW^T \\ &\quad \Rightarrow FWW^T(WW^T)^{-1} = HW^T(WW^T)^{-1} \\ &\quad \Rightarrow F = HW^T(WW^T)^{-1} \end{aligned} \tag{7}$$

where $W^T(WW^T)^{-1}$ is the pseudo inverse of the latent semantic-inquiry matrix W.

Therefore, for a given incoming inquiry g, we can apply both projections to derive the informational goal:

$$FU^T g = Fw = h \tag{8}$$

where g is the vector of weights for all terms in the test inquiry; and w is the vector of weights for each latent semantic corresponding to the test inquiry.

An informational goal a'_n will be assigned the automatically derived informational goal for inquiry n where $a'_n = \arg\max_a\{h_{a,n}\}$. Overall accuracy is defined in terms of the fraction of inquiries in the test set with correctly inferred informational goals.

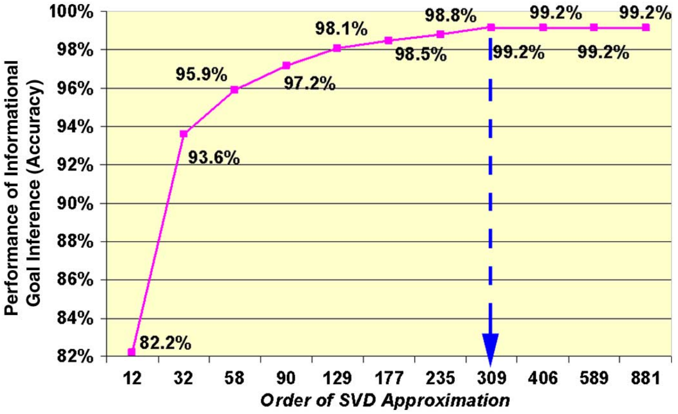


Fig. 3. A plot of the accuracy of singular values retained against the dimensionality of the SVD approximation.

In order to evaluate the effect of imperfect speech and pen recognition transcriptions on informational goal inference, we will develop projections (i.e. FU^T) based on both perfect (i.e. manual) and imperfect (i.e. speech recognition) transcriptions of the training inquiries. We will also test with both perfect and imperfect transcriptions of the testing inquiries G. These four combinations can be summarized as:

Transcription Type used on Deriving Projections	Transcription Type in the Test Set	Notation of the Goal-Inquiry Matrix
Perfect (i.e. $F_1U_1^T$)	Perfect (i.e. G_1)	H_{11}
Perfect (i.e. $F_1U_1^T$)	Imperfect (i.e. G_2)	H_{12}
Imperfect (i.e. $F_2U_2^T$)	Perfect (i.e. G_1)	H_{21}
Imperfect (i.e. $F_2U_2^T$)	Imperfect (i.e. G_2)	H_{22}

A. Implementation of Semantic Inference

Recall that the non-negative matrix G (see Equation (1) has dimensionality 881×999. We use Equation (3) (i.e. $G = USV^T$) to project the term-inquiry space onto an R-dimensional latent semantic space. Based on the latent semantic space, we may reduce and reconstruct the term-inquiry space using:

$$G \approx \hat{G} = U\hat{S}V^T. \tag{9}$$

where $\hat{S}$ is the reduced diagonal matrix of singular values with optimized order of SVD approximation (i.e. R').

In order to collapse the terms that are "semantically similar," we choose $R' < R$. The smaller the value of R,' the more pronounced is the reduction of semantic redundancy in the latent semantic space. We need to find an "optimal" choice of R' that minimizes the distortion between the re-constructed space $\hat{G}$ and the original space G during the implementation of Equation (3) in the training procedure. We plan to optimize R' through empirical analysis of the latent space.

We choose the possible values of R' with reference to the percentage of the cumulative sum of retained singular values over the maximum at $R' = R = 881$, searching at intervals of 10%. We then perform informational goal inference on the multimodal inputs in the training set at different values of R' (see Fig. 3). The performance of informational goal inference increases with R.' The rate of increase slowed down as R' becomes higher, reaching saturation at approximately $R' = 309$.

TABLE VI
ACCURACY OF INFORMATIONAL GOAL INFERENCE USING VSM FOR DIFFERENT COMBINATIONS OF PERFECT/IMPERFECT INFORMATIONAL GOAL VECTORS AND TEST INQUIRIES

Informational goal vectors	Test inquiries	Training Set	Test Set
Perfect	Perfect	84.5%	82.5%
Perfect	Imperfect	19.2%	17.8%
Imperfect	Perfect	74.6%	73.8%
Imperfect	Imperfect	57.9%	46.0%

We then perform informational goal inference in the training set at finer resolution, among values of R' between 235 and 309. The inference performance reaches saturation at $R' = 263$, which implies a reduction of semantic redundancy of 70% with respect to the original space of $R = 881$.

VII. PERFORMANCE BASELINE USING THE VECTOR-SPACE MODEL

To demonstrate the efficacy of the LSA approach, we choose a reference baseline using the vector-space model (VSM) [44] for semantic inference.[6] We sum the weights of every multimodal or lexical term (obtained using Equation (2) across the training inquiries of each informational goal and normalize the sum with the total number of inquiries for that informational goal. We then create a vector j_a of the normalized weights, where a denotes the informational goal. For an incoming test inquiry, we create a vector g using the same method. The similarity between the vectors g and j_a is calculated as the inner product, denoted by Z_a in Equation (10) [44], [45], normalizing for the length of the vectors.

$$Z_a = \frac{j_a \cdot g}{\|j_a\| \, \|g\|} \tag{10}$$

where Z_a denotes the similarity between vectors j_a and g; j_a is the weight for all terms in the ath informational goal; and g is the weight for all terms in the test inquiry.

The input expression n is assigned to the informational goal a_n^* which has the maximum similarity score, as shown in Equation (11).

$$a_n^* = \arg\max_a \{Z_a\} \tag{11}$$

Experiments show that VSM can correctly infer informational goals for 84.5% and 82.5% of the inquiries in training and test sets from *perfect* transcriptions respectively. We then apply the VSM to different combinations of perfect and imperfect transcriptions. Detailed results are shown in Table VI.

VIII. SEMANTIC INFERENCE WITH LATENT PROJECTION DERIVED FROM PERFECT TRANSCRIPTIONS

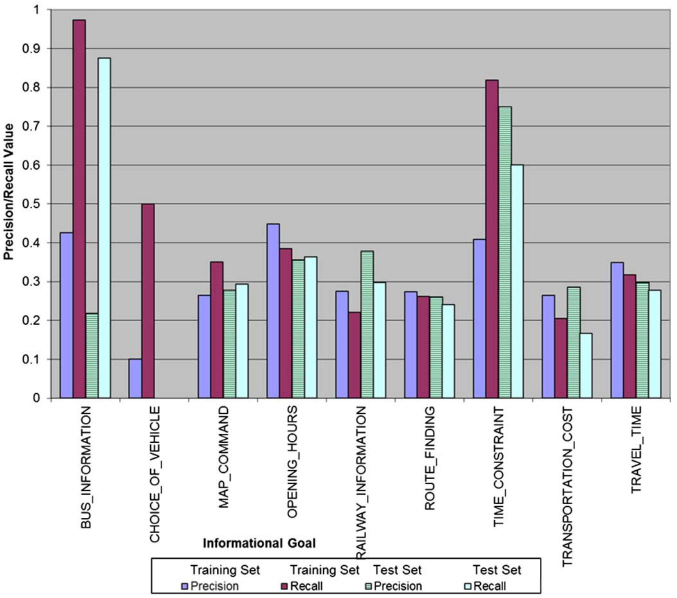


Fig. 4. Performance of informational goal inference using the perfect projection model and *imperfect* test set transcription. Results are based on the latent space with $R' = 263$ dimensions.

Recall that we have developed the projections (i.e. FU^T) based on manual transcriptions, which is referred as the "perfect projection model" below. We use the perfect projection model $F_1 U_1^T$ to project *perfect* test inquiries G_1 (i.e. manually transcribed test inquiries) to the space of informational goals, i.e. $F_1 U_1^T G_1 = H_{11}$. Overall accuracy in informational goal inference for the training and test sets are 99.2% and 98.6% respectively. The test set lacks inquiries that fall under the informational goal of CHOICE_OF_VEHICLE.

We then apply the projection derived from perfect speech and pen transcriptions to *imperfectly transcribed* test inquiries for informational goal inference, i.e. $F_1 U_1^T G_2 = H_{12}$. Overall accuracy in informational goal inference for the training and test sets drops to 32.2% and 28.4% respectively. Precision and recall[7] of each informational goal are shown in Fig. 4. Informational goal of inquiries that fall under the informational goal of TIME_CONSTRAINT are all inferred incorrectly. The poor performance is due to the great mismatch in the term lists between perfect and imperfect transcriptions. There are only 881 terms (including multimodal and lexical terms) in the term list of perfect (i.e. manual) transcriptions, as compared with 1,901 terms in the term list of imperfect (i.e. automatic) transcriptions (i.e. test inquiries). Many terms in the test set are not covered by the perfect projection model. For example, 大街 (pronounced as /*da jie*/, meaning "*main street*") is misrecognized as 大世界 (/*da shi jie*/ meaning "*big world*"), 大姐 (/*da jie*/ meaning "*elder sister*"), 大家 (/*da jia*/ meaning "*we*"), 大概(/*da gai*/ meaning "*about*"), 大寫 (/*da xie*/ meaning "*capital letter*"), 大學(/*da xue*/ meaning "*university*"), 大戰(/*da zhan*/ meaning "*war*"), etc. and the misrecognition outputs are not covered by the perfect projection model.

[6] We have experimented with the VSM model as well as single- and multi-class support vector machines (SVM) and are reporting on the higher reference baseline.

[7] Precision is the percentage of inquiries in the test set correctly inferred as informational goal a, out of all the inquiries which are inferred as informational goal a. Recall is the percentage of inquiries in the test set correctly inferred as informational goal a, out of all the inquiries which truly belongs to informational goal a.

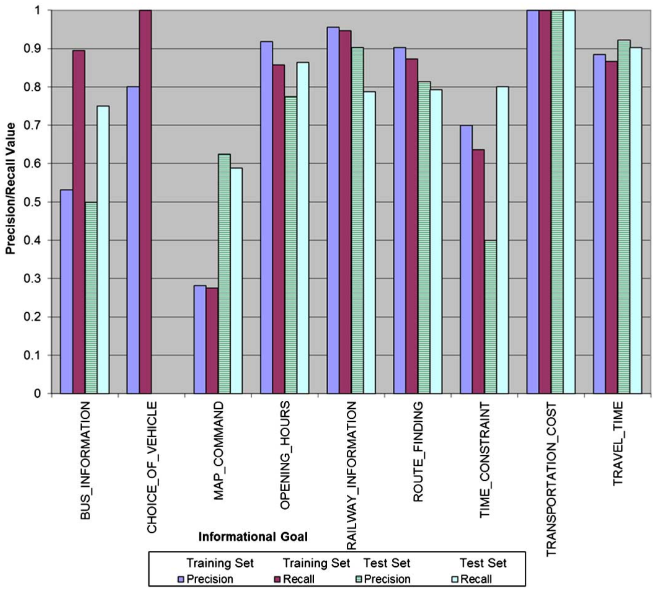


Fig. 5. Performance of informational goal inference using the imperfect projection model and *imperfect* test set transcriptions. Results are based on the latent space with $R' = 650$ dimensions.

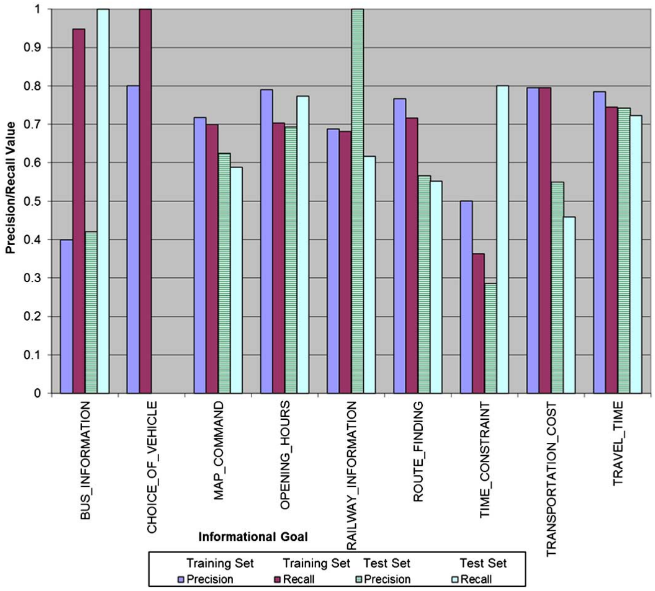


Fig. 6. Performance of informational goal inference using the imperfect projection model and *imperfect* test set transcriptions. Results are based on the latent space with $R' = 650$ dimensions.

IX. Semantic Inference with Latent Projection Derived from Imperfect Transcriptions

The previous section presented results where perfect transcriptions of the training set are used to derive the latent projection using LSA. This section presents the use of imperfect transcriptions of the training set to derive the latent projection model, hereafter referred as the "imperfect projection model." There are a total of 797 multimodal and 1,104 lexical terms in the recognition output of the training set. The number of multimodal and lexical terms increased from 567 and 314 respectively for manual transcriptions to 797 and 1,104 respectively for automatic transcriptions (i.e. recognition outputs). The increase is mainly due to misrecognition of Chinese measure words while the increase in lexical terms is due to character recognition errors and confusion with English words. For example, 遊覽 (*/you lan/* meaning "*tour*") is confused with u-盤 (*/u-pan/* meaning "*USB drive*"), "you," "yahoo," "ultra," "soul," etc. The total number of terms sum to $M = 1901$. Hence the non-negative matrix G (see Equation (1)) has increased dimensionality at 1901×999. As described in Section VI.1, we apply SVD and follow similar procedures to reduce dimensionality to $R' = 650$. This corresponds to a dimension reduction of 66% compared with the original space of $R = 1901$.

We apply the imperfect projection model to perfect test set transcriptions, i.e. $F_2 U_2^T G_1 = H_{21}$. Overall accuracy in informational goal inference for the training and test sets are 85.8% and 83.7% respectively. Detailed results are shown in Fig. 5. Next, we apply the imperfect projection model on the imperfectly transcribed test set, i.e. $F_2 U_2^T G_2 = H_{22}$. Overall accuracy in informational goal inference for the training and test sets are 73.0% and 64.3% respectively. Detailed results are shown in Fig. 6.

X. Analysis of the Latent Semantic Space for Informational Goal Inference

Comparison between latent projection models developed using perfect and imperfect transcriptions shows similar levels of dimension reduction (i.e. 70% and 66% respectively). The summary of performance accuracies in informational goal inferences (see Table VII) shows that when the perfect projection model is used, there is a significant drop in performance as we migrate from perfect to imperfect test set transcriptions, i.e. from 99.2% to 32.2% and 98.6% to 28.4% in training and test sets respectively. This is because only 16.7% of the terms in the imperfect test inquiries are covered in the perfect projection model, which leads to a serious out-of-vocabulary problem in informational goal inference. Corresponding values based on the imperfect projection model show a performance drop from 85.8% to 73% for the training set, and from 83.7% to 64.3% for the test set. Analysis shows that around half (52.5%) of the terms in the perfect test inquiries are covered in the imperfect projection model. The imperfect projection model shows greater robustness in informational goal inference. Comparison between the statistics in Tables VI and VII also shows that LSA always performs significantly better than VSM in informational goal inference (the difference ranges from 9.9% to 18.3%). This suggests that reduction into the latent space can capture better the relevant semantics and contextual information for informational goal inference. We will elaborate on this in the following.

TABLE VII
Performance of Informational Goal Inference Using Different Combinations of Perfect/Imperfect Projection Models and Test Inquiries

Projections	Test inquiries	Training Set	Test Set
Perfect	Perfect	99.2%	98.6%
Perfect	Imperfect	32.2%	28.4%
Imperfect	Perfect	85.8%	83.7%
Imperfect	Imperfect	73%	64.3%

A. Sub-categorization of Informational Goals

Matrix F gives the weight of each informational goal with each latent semantic category. Hence, we can select the latent

TABLE VIII
EXAMPLES OF INQUIRIES THAT BELONG TO THE LATENT SEMANTIC CATEGORIES 13 AND 19 FROM THE TRAINING SET FOR THE INFORMATIONAL GOAL BUS_INFORMATION. TRANSLATIONS ARE ITALICIZED

***r*=13 corresponds to BUS_INFORMATION along a street**
經過 <這條大街|STROKE|SEQ> 的 所有 公交 線路 是 哪些 "*What are the bus routes that pass through <this street|STROKE|SEQ>*"
***r*=19 corresponds to BUS_INFORMATION within an area**
告訴 我 所有 在 <這個範圍|CIRCLE|SIM> 行走 的 公交 路線 "*Tell me what are the bus routes that pass through <this area|CIRCLE|SIM>*"

semantic categories with high weights to be representatives of a goal. Furthermore, matrix W gives the weights relating latent semantic categories and the inquiries. Hence, we can select inquiries with high weights to be representatives of a latent semantic category. We examine the matrices F and W obtained from the *perfect* transcriptions (i.e. F_1 and $U_1^T G_1$) and find that the semantic inference approach is able to sub-categorize the informational goals into logical sub-types and capture their key terms. For example, the informational goal BUS_INFORMATION contains two latent semantic categories: (i) bus information along a street ($r = 13$) and (ii) bus information within an area ($r = 19$). Table VIII shows some examples. We note a couple of key usage patterns in this sub-category, e.g. < 這條大街|STROKE|SEQ> (or "$< this\ street|STROKE|SEQ >$") is a common way of referring to a street, with the uttered phrase "*this street*" occurring sequentially with a stroke. Another example is <這個範圍|CIRCLE|SIM> (or "$< this\ area|CIRCLE|SIM >$") where circling gesture occurs simultaneously with the uttered key phrase of "*this area*."

Another example is the informational goal OPENING_HOURS, which contains six latent semantic categories: (i) opening hours of one location ($r = 11$); (ii) opening hours of a single or multiple locations using ellipsis ($r = 46$); (iii) opening hours of multiple locations using multiple singular SLRs ($r = 7$ and 29); (iv) opening hours of multiple locations using one aggregated SLR ($r = 9$); and (v) opening hours of multiple locations using one plural SLR ($r = 12$). Table IX shows some examples. We also note some examples of key usage patterns in this Table. In particular, inquiries about OPENING_HOURS involve ellipsis, denoted by $< \emptyset|\mathrm{POINT}|\emptyset >$.

Such sub-categorization based on latent semantics is potentially advantageous because finer semantics categorization can enhance understanding and will facilitate automatic generation of system responses.

B. Capturing Key Terms for Informational Goals

We examine the term weights based on the *perfect* and *imperfect* transcriptions in the latent semantic space to identify key terms that are indicative of each informational goal. Figs. 7 to 9 are the plots of term weights from informational goal-term matrix F against terms (both lexical and multimodal terms), for the informational goals ROUTE_FINDING and TIME_CONSTRAINT respectively.

Comparison between Figs. 8 and 9 shows that we can capture similar key terms from perfect and imperfect transcriptions

TABLE IX
EXAMPLES OF INQUIRIES THAT BELONG TO THE SIX LATENT SEMANTIC CATEGORIES FROM THE TRAINING SET FOR INFORMATIONAL GOAL OPENING_HOURS. TRANSLATIONS ARE ITALICIZED

***r*=11 corresponds to OPENING_HOURS of one location**
我 想 知道 <這裡|POINT|SIM> 的 開放時間 "*I want to know the opening hours of <here|POINT|SIM>*"
***r*=46 corresponds to OPENING_HOURS of single or multiple locations using ellipsis**
<Ø|POINT|Ø> 開放時間 "*<Ø|POINT|Ø> opening hours*"
***r*=7 and 29 correspond to OPENING_HOURS of multiple locations using multiple singular SLRs**
我 想 知道 <這個市場|POINT|SIM> <這個廣場|POINT|SIM> <這個購物中心|POINT|SIM> 的 營運時間 "*I would like to know the opening hours of <this plaza|POINT|SIM>, <this plaza|POINT|SIM> and <this shopping center|POINT|SIM>*"
***r*=9 corresponds to OPENING_HOURS of multiple locations using aggregated SLR**
勞駕 你 告訴 我 <這三個地方|MULTI-POINT|SIM> 的 營業時間 "*Please tell me the opening hours of <these three places|MULTI-POINT|SIM>*"
***r*=12 corresponds to OPENING_HOURS of multiple locations using one plural SLR**
請問 <這幾個地方|MULTI-CIRCLE|SEQ> 的 開放時間 是 從 幾點 到 幾點 "*The opening hours of <these locations|MULTI-CIRCLE|SEQ> are from when to when*"

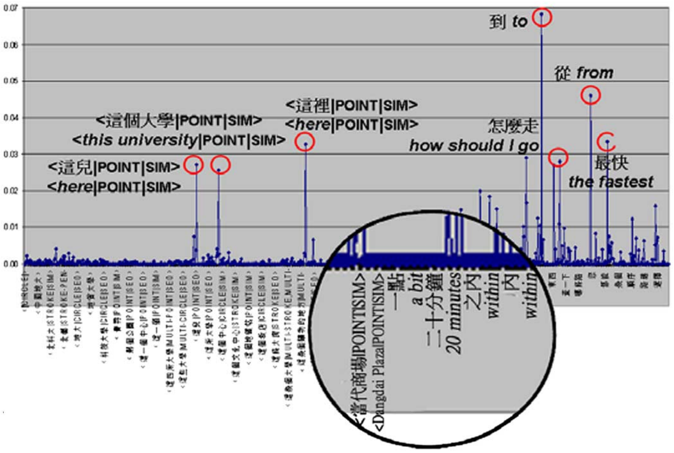


Fig. 7. A plot of term weights from matrix F_1 against 881 terms from *perfect* transcriptions of inquiries with the informational goal ROUTE_FINDING. Some of the multimodal and lexical terms on the x-axis are magnified. The first magnified term is multimodal $< Dangdai\ Plaza|POINT|SIM >$ while the remaining are lexical terms.

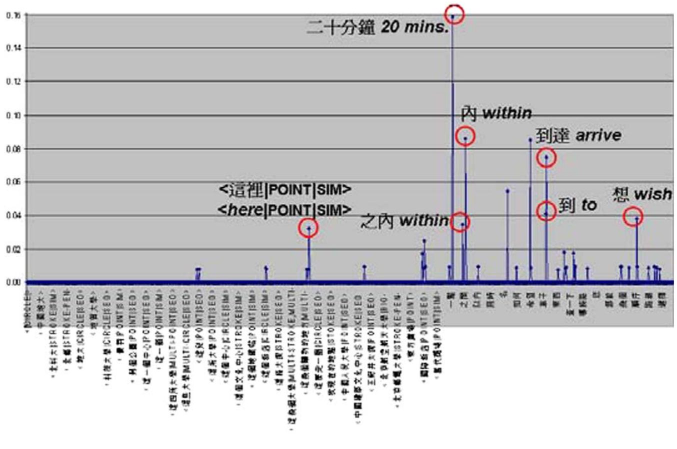


Fig. 8. A plot of term weights from matrix F_1 against 881 terms from *perfect* transcriptions of inquiries with the informational goal TIME_CONSTRAINT. The x-axis is the same as Fig. 7, covering all terms from the perfect transcriptions of the training set.

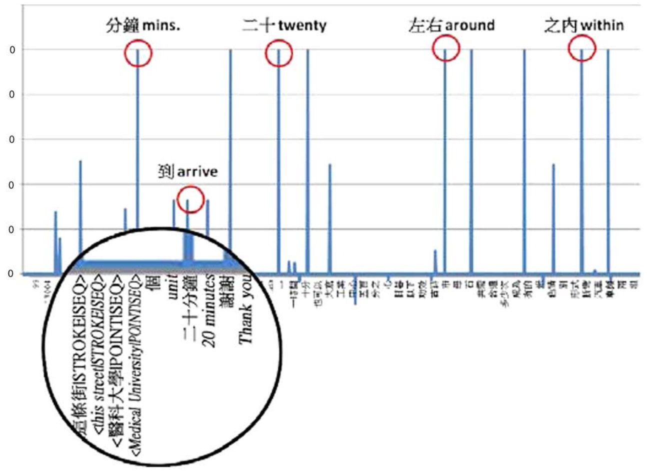


Fig. 9. A plot of term weights from matrix F_2 against 1,901 terms from *imperfect* transcriptions of inquiries with the informational goal TIME_CONSTRAINT. Some of the multimodal and lexical terms on the x-axis are magnified. Note that the x-axis here is different from Fig. 10. This is because this plot is based on imperfect transcription of the training set.

for the same informational goal TIME_CONSTRAINT. For example, "到達," "到" (meaning "*arrive*"), "之內," "內" (meaning "*within*"), "二十分鐘," "二十" and "分鐘" (meaning "*twenty minutes*"). In the last example, we also observe the ambiguity in Chinese word segmentation here.

XI. Conclusions and Future Work

This paper describes our work in semantic interpretation of a "multimodal language" with speech and gestures using latent semantic analysis (LSA). We apply speech and pen recognition to transcribe multimodal user inputs in a navigation domain. We parse for spoken locative references (SLRs) in the speech modality and attempt to partially interpret the events in the pen modality. We then apply Viterbi alignment that applies semantic and temporal constraints to generate associations between the SLRs and the pen input. In the LSA approach, we use a non-negative term-inquiry matrix to capture the associations between terms (both lexical and multimodal terms) and inquiries. Factorization of the term-inquiry matrix using singular value decomposition (SVD) deduces the associations between terms and inquiries through a latent semantic space with reduced dimensionality. We project the latent semantic space into the space of informational goals through a matrix derived from training set. An input multimodal inquiry can be projected into the latent semantic space and then into the informational goal space. This gives rise to a vector with which we can use the highest weighting element to select the inferred informational goal. We experimented with projections derived from both perfect (i.e. via manual labeling) and imperfect (i.e. via automatic recognition) transcriptions of the training set. These are referred as the perfect and imperfect projection models respectively. These models are applied to the perfect and imperfect transcriptions of the test set. The LSA approach attains comparable degrees of dimension reduction for both perfect and imperfect transcriptions. In terms of informational goal inference, the perfect projection model achieves 98.6% accuracy on a perfectly transcribed test set. This degrades to 28.4% when imperfect transcriptions are used. The imperfect projection model achieves 83.7% and 64.3% accuracies respectively based on perfect and imperfect test set transcriptions and hence exhibits greater robustness in informational goal inference. Comparison with the baseline performance using the VSM approach shows that the LSA always performs significantly better (at least 9.9% absolute) in informational goal inference. This suggests that reduction into the latent space can capture better the relevant semantics and contextual information for informational goal inference.

Acknowledgment

We would like to thank the anonymous reviewers for their helpful comments.

References

[1] S. Oviatt, A. DeAngeli, and K. Kuhn, "Integration and synchronization of input modes during multimodal human-computer interaction," in *Proc. SIGCHI Conf. Human Factors Comput. Syst.*, Atlanta, GA, USA, Mar. 22–27, 1997, pp. 415–422.

[2] M. Johnston, P. Cohen, D. McGee, S. Oviatt, J. Pittman, and I. Smith, "Unification-based multimodal integration," in *Proc. 35th Annu. Meeting Assoc. Comput. Linguist. 8th Conf. Eur. Chapter Assoc. Comput. Linguist. (COLING-ACL)*, Madrid, Spain, 1997, pp. 281–288.

[3] S. Wang, "A multimodal galaxy-based geographic system," S.M. thesis, Mass. Inst. Technol., Cambridge, MA, USA, 2003.

[4] M. Johnston and S. Bangalore, "Multimodal applications from mobile to kiosk," in *Proc. W3C Workshop Multimodal Interaction*, Sophia Antipolis, France, Jul. 2004, pp. 19–20.

[5] R. Malaka, J. Haeussler, and H. Aras, "SmartKom mobile–intelligent ubiquitous user interaction," in *Proc. 9th Int. Conf. Intell. User Interfaces*, Madeira, Portugal, Jan. 13–16, 2004, pp. 310–312.

[6] K. Wang, "From multimodal to natural interactions," in *Proc. W3C Workshop Multimodal Interaction*, Sophia Antipolis, France, Jul. 2004, pp. 19–20.

[7] J. Y. Chai, P. Hong, M. X. Zhou, and Z. Prasov, "Optimization in multimodal interpretation," in *Proc. 42nd Annu. Meeting Assoc. Comput. Linguist.*, Barcelona, Spain, Jul. 21–26, 2004, pp. 1–8.

[8] S. Qu and J. Chai, "Salience modeling based on non-verbal modalities for spoken language understanding," in *Proc. 8th Int. Conf. Multimodal Interfaces*, Nov. 2–4, 2006, pp. 193–200.

[9] L. J. Wang, T. Hu, P. Liu, and F. Soong, "Efficient handwriting correction of speech recognition errors with template constrained posterior (TCP)," in *Proc. Interspeech*, Brisbane, Australia, Sep. 22–26, 2008, pp. 2659–2662.

[10] K. Shinoda, Y. Watanabe, K. Iwata, Y. Liang, R. Nakagawa, and S. Furui, "Semi-synchronous speech and pen input for mobile user interfaces," *Speech Commun.*, vol. 53, pp. 283–291, 2011.

[11] L. Nigay and J. Coutaz, "A generic platform for addressing the multimodal challenge," in *Proc. ACM Conf. Human Factors Comput. Syst. (CHI)*, May 7–11, 1995, pp. 98–105.

[12] E. Filisko and S. Seneff, "A context resolution server for the galaxy conversational systems," in *Proc. 8th Eur. Conf. Speech Commun. Technol. (Eurospeech-Interspeech 2003)*, Geneva, Switzerland, Sep. 1–4, 2003, pp. 197–200.

[13] M. Johnston, "Unification-based multimodal parsing," in *Proc. 36th Annu. Meeting Assoc. Comput. Linguist. 17th Int. Conf. Comput. Linguist. (COLING-ACL)*, Montreal, QC, Canada, 1998, vol. 1, pp. 624–630.

[14] L. Wu, S. Oviatt, and P. Cohen, "Multimodal integration - a statistical view," *IEEE Trans. Multimedia*, vol. 1, no. 4, pp. 334–341, Dec. 1999.

[15] W. Wahlster, *SmartKom: Foundations of Multimodal Dialogue Systems*. Berlin/Heidelberg, Germany: Springer, Sep. 2006.

[16] M. Johnston and S. Bangalore, "Finite-state multimodal parsing and understanding," in *Proc. 18th Int. Conf. Comput. Linguistics (COLING)*, Jul.-Aug. 31–4, 2000, vol. 1, pp. 369–375.

[17] M. Johnston and S. Bangalore, "Finite-state multimodal integration and understanding," *J. Natural Lang. Eng.*, vol. 11, no. 2, pp. 159–187, 2005.

[18] S. Bangalore and M. Johnston, "Robust understanding in multimodal interfaces," *J. Comput. Linguist.*, vol. 35, no. 3, pp. 345–397, Sep. 2009.

[19] J. Y. Chai, P. Hong, and M. X. Zhou, "A probabilistic approach to reference resolution in multimodal user interfaces," in *Proc. 9th Int. Conf. Intell. User Interfaces (IUI)*, Funchal, Portugal, Jan. 13–16, 2004, pp. 70–77.

[20] T. M. Sezgin, I. Davies, and P. Robinson, "Multimodal inference for driver-vehicle interaction," in *Proc. 11th Int. Conf. Multimodal Interfaces, (ICMI-MLMI)*, Cambridge, MA, USA, Nov. 2–6, 2009, pp. 193–198.

[21] S. Qu and J. Y. Chai, "Beyond attention: The role of deictic gesture in intention recognition in multimodal conversational interfaces," in *Proc. Int. Conf. Intell. User Interfaces (IUI)*, Maspalomas, Spain, Jan. 13–16, 2008, pp. 237–246.

[22] J. Bellegarda, "Statistical language model adaptation: Review and perspectives," *Speech Commun.*, vol. 42, no. 1, pp. 93–108, 2004.

[23] J. R. Bellegarda, "Latent semantic mapping: Principles and applications," *Synth. Lectures Speech Audio Process.*, vol. 3, no. 1, pp. 1–101, 2007.

[24] S. Deerwester, S. Dumais, T. Landauer, G. Furnas, and R. Harshman, "Indexing by latent semantic analysis," *J. Amer. Soc. Inf. Sci.*, vol. 41, pp. 391–407, 1990.

[25] S. Dumais, "Using LSI for information filtering: TREC-3 experiments," in *Third Text Retrieval Conf. (TREC3)*, D. Harman, Ed. Gaithersburg, MD, USA: National Institute of Standards and Technology Special Publication, 1995.

[26] M. Berry, S. Dumais, and G. W. O'Brien, "Using linear algebra for intelligent information retrieval," *SIAM Rev.*, vol. 37, pp. 573–595, 1995.

[27] T. Landauer and S. Dumais, "A solution to plato's problem: The latent semantic analysis theory of acquisition, Induction and representation of knowledge," *Psychological Review*, vol. 104, pp. 211–240, 1997.

[28] J. Bellegarda, J. Butzberger, Y. Chow, N. Coccaro, and D. Naik, "A novel word clustering algorithm based on latent semantic analysis," in *Proc. Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, 1996, vol. 1, pp. 172–175.

[29] Y. Gotoh and S. Renals, "Document space models using latent semantic analysis," in *Proc. 5th Eur. Conf. Speech Commun. Technol. (Eurospeech)*, Rhodes, Greece, Sep. 1997, pp. 22–25.

[30] D. Gildea and T. Hofmann, "Topic-based Language Models using EM," in *Proc. 6th Eur. Conf. Speech Commun. Technol.*, Budapest, Hungary, Sep. 1999, pp. 5–9.

[31] J. R. Bellegarda, "Exploiting latent semantic information in statistical language modeling," *Proc. IEEE Special Autom. Speech Recogn. Understand. Spoken Lang.*, vol. 88, no. 8, pp. 1279–1296, Aug. 2000.

[32] P. W. Foltz and S. T. Dumais, "Personalized information delivery: An analysis of information filtering methods," *Commun. ACM*, vol. 35, pp. 51–60, 1992.

[33] J. R. Bellegarda, D. Naik, and K. E. A. Silverman, "Automatic junk E-mail filtering based on latent content," in *Proc. Automat. Speech Recognition Understanding Workshop*, St. Thomas, U. S. Virgin Islands, Dec. 2003, pp. 465–470.

[34] S. T. Dumais and J. Nielsen, "Automating the assignment of submitted manuscripts to reviewers," in *Proc. ACM SIGIR'92 15th Int. Conf. Research Develop. Inf. Retrieval*, Copenhagen, Denmark, Jun. 21–24, 1992, pp. 233–244.

[35] R. Serafin and B. D. Eugenio, "FLSA: Extending latent semantic analysis with features for dialogue act classification," in *Proc. ACL*, 2004, pp. 692–699.

[36] J. R. Bellegarda and K. E. A. Silverman, "Natural language spoken interface control using data-driven semantic inference," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 11, no. 3, pp. 267–277, May 2003.

[37] P. Y. Hui and H. Meng, "Cross-modality semantic integration with hypothesis rescoring for robust interpretation of multimodal user interactions," *IEEE Trans. Audio, Speech Lang. Process.*, vol. 17, no. 3, pp. 486–500, Mar. 2009.

[38] "WordNet," [Online]. Available: http://wordnet.princeton.edu/

[39] B. Xiao, C. Girand, and S. Oviatt, "Multimodal integration patterns in children," in *Proc. 7th Int. Conf. Spoken Lang. Process. (ICSLP)*, Denver, CO, USA, Sep. 16–20, 2002, pp. 629–632.

[40] B. Xiao, R. Lunsford, R. Coulston, M. Wesson, and S. Oviatt, "Modeling multimodal integration patterns and performance in seniors: Towards adaptive processing of individual differences," in *Proc. 5th Int. Conf. Multimodal Interfaces (ICMI)*, Vancouver, BC, Canada, Nov. 5–7, 2003, pp. 265–272.

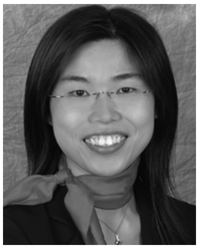
[41] P. F. Brown, S. A. D. Pietra, V. J. D. Pietra, and R. L. Mercer, "The mathematics of machine translation: Parameter estimation," *Comput. Linguist.*, vol. 19, no. 2, pp. 263–311, 1993.

[42] J. R. Bellegarda, "Exploiting latent semantic information in statistical language modeling," *Proc. IEEE*, vol. 88, no. 8, pp. 1279–1296, Aug. 2000.

[43] G. Salton and C. Buckley, "Term-weighting approaches in automatic text retrieval," *An Int. J. Inf. Proc. Manag.*, vol. 24, no. 5, pp. 513–523, 1998.

[44] G. Salton and M. McGill, *Int. Modern Inf. Retrieval*. New York, NY, USA: McGraw-Hall, 1983.

[45] C. J. van Rijsbergen, *Information Retrieval*. London, U.K.: Butterworths, 1979.



Pui-Yu Hui received her Bachelor, M. Phil. and Ph.D. degrees, all in Systems Engineering and Engineering Management from The Chinese University of Hong Kong. She was a Postdoctoral Fellow of the same department. Her research interests include spoken document retrieval, multimodal input integration and understanding; and computer-aided learning system.

Helen M. Meng received the S.B., S.M., and Ph.D. degrees, all in electrical engineering, from the Massachusetts Institute of Technology, Cambridge. She has been a Research Scientist at the MIT Spoken Language Systems Group, where she worked on multilingual conversational systems. She joined The Chinese University of Hong Kong in 1998, where she is currently Professor and Chairman of the Department of Systems Engineering and Engineering Management. She was also the past Associate Dean 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-Microsoft Key Laboratory in 2008, and serves as Co-Director. She is also Co-Director of the Tsinghua-CUHK Joint Research Center for Media Sciences, Technologies and Systems. Helen's research interest is in the area of human-computer interaction via multimodal and multilingual spoken language systems, as well as translingual speech retrieval technologies. She has been elected IEEE Fellow in 2013 and Editor-in-Chief of the IEEE TRANSACTIONS ON AUDIO, SPEECH AND LANGUAGE PROCESSING. She is also an elected board member of the International Speech Communication Association.