

# Complementarity and Redundancy in Multimodal User Inputs with Speech and Pen Gestures

Pui-Yu Hui, Zhengyu Zhou and Helen Meng

Human-Computer Communications Laboratory

Department of Systems Engineering and Engineering Management

The Chinese University of Hong Kong

Shatin, Hong Kong SAR of China

{pyhui, zyzhou, hmmeng}@se.cuhk.edu.hk

## Abstract

We present a comparative analysis of multi-modal user inputs with speech and pen gestures, together with their semantically equivalent uni-modal (speech only) counterparts. The multimodal interactions are derived from a corpus collected with a Pocket PC emulator in the context of navigation around Beijing. We devise a cross-modality integration methodology that interprets a multi-modal input and paraphrases it as a semantically equivalent, uni-modal input. Thus we generate parallel multi-modal (MM) and uni-modal (UM) corpora for comparative study. Empirical analysis based on class trigram perplexities shows two categories of data: ( $PP_{MM} = PP_{UM}$ ) and ( $PP_{MM} < PP_{UM}$ ). The former involves complementarity across modalities in expressing the user's intent, including occurrences of ellipses. The latter involves redundancy, which will be useful for handling recognition errors by exploring mutual reinforcements. We present explanatory examples of data in these two categories.

**Index terms:** multi-modal input, spoken input, pen gesture, joint interpretation, human-computer interaction, perplexity

## 1. Introduction

This paper presents a comparative analysis of multimodal (speech and pen gestures) user inputs with their semantically equivalent unimodal (speech only) counterparts, in order to gain an empirical understanding of the inter-relations between the speech and pen modalities. Increasing use of mobile handheld devices for information access in our daily lives has led to the growing prominence of multimodal user interfaces (MUI). Users may either use speech as a hands-free modality or switch to pen gestures in noisy ambient conditions. Additionally, users may use both modalities in coordination for enhanced expressive power, especially in the communication of complex semantics in a succinct form [1]. For example, the unimodal, spoken inquiry:

*What is the name of the street that is five blocks south of the Yonghegong, intersection and east of the China National Museum of Fine Arts?*

may be paraphrased multimodally with substantial simplification, as:

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


Interpretation of multimodal user expressions is gaining increasing interest from our research community [2, 3, 4, 5, 6]. The interpretation framework must capture semantic relationships across modalities, such as the CARE (complementarity, assignment, redundancy and equivalence) properties as identified in [2]. Complementary and redundant relations in input modes are further described in [7]. This paper leverages previous research and attempts to form an empirical, organizational view of multimodal integration patterns. Our long-term goal is to develop techniques for automatic semantic interpretation of multimodal user input as a front-end extension to unimodal spoken dialog systems. We begin with a comparative analysis between the multimodal inputs and their unimodal counterparts. We collected a

multimodal corpus based on user interactions with a Pocket PC (PPC) emulator to seek navigational information about the Beijing area [5]. We have also devised a cross-modality integration model that accepts multimodal user inputs and generates semantically equivalent unimodal paraphrases. We trained a trigram language model using pooled multimodal and unimodal data in a training set. We computed test set perplexities of disjoint, parallel test sets with multimodal and unimodal inputs respectively. Comparison of perplexities enables categorization into subsets for further analysis. Details of our approach are presented in the following.

## 2. The Multi-modal Corpus

Our experimental corpus is collected with a Pocket PC (PPC) emulator with which the user interacts in order to obtain navigational information about Beijing. This information domain involves references to maps and the communication of spatial semantics. The scope of the domain involves 6 maps (that fit the PPC screen-size) covering 5 districts and 930 locations with positional coordinates. There is a variety of location types (such as parks, streets and universities), as well as communicative goals on the part of the users (such as bus fares, route-finding and travel time). Data collection involves 21 subjects from a speech research group. Each subject is asked to formulate set of input inquiries or requests based on a navigational task. The inputs may involve a spoken command (e.g. for map rendering), or a spoken question that references up to a maximum of six locations. The subjects are free to refer to the locative semantics either unimodally (with speech only) or multimodally (with speech and pen gestures). A typical user input may contain up to six spoken locative references and/or pen gestures. We collected 1,386 user inputs in total. Among these, 320 are unimodal and 1066 are multimodal inputs. We used approximately 70% of the multimodal utterances as training data (for analysis and parameter selection) and the remaining 30% as testing data. Recorded speech is in Chinese and has been endpointed and hand-transcribed (i.e. perfect transcriptions). The utterances cover 519 Chinese lexical entries and range between one to 28 words in utterance lengths. The multimodal inputs have 2,570 pen gestures in total, including pointing, circling and strokes. Each pen gesture is recorded with a time-stamp and relevant (x,y) coordinate(s), e.g. the pen-down and pen-up actions in a stroke. There are also spurious gestures that were captured during data collection but these are filtered out automatically. An example of a multimodal input is:

Table 1. An example of a multimodal user input with speech and pen gestures.

Speech: 我從 這裡 要到這四個大學一共需要多少時間?
Pen: 
Translation of the spoken utterance: "I want to go from <u>here</u> to <u>these four universities</u> . How much time will it take?" (Input pen gestures include a point and a circle.)

### 3. Cross-Modality Integration

Each modality in a multimodal input abstracts the user’s message differently into a sequence of input events, e.g. spoken keywords/keyphrases or pen gestures. Each carries semantic meaning but may contain ambiguity. For example, the user may refer to a location directly by its name or abbreviation, e.g. “CUG” for “中國地質大學” (China University of Geosciences) and these direct references have little ambiguity. However, there are also indirect spoken references, or spoken deictic expressions, e.g. “這裡” (here), “這些地方” (these places) or “這四所大學” (these four universities). As can be seen, these indirect deictic expressions may carry numeric features or location type information. They may also be semantically ambiguous. Similarly, for pen-based input, a pointing gesture may refer to the map’s coordinates (e.g. “zoom in here” <point>) or a location (e.g. a landmark); a circling gesture may refer to a single location, a group of locations or a region; and a stroke may refer to one more locations, a path or a demarcation. Hence, pen gestures may also contain considerable ambiguity.

We devised a cross-modality integration framework that accepts a multimodal input expression and generate a unimodal paraphrase. The speech modality is first parsed for spoken locative reference expressions. For each expression, our framework generates a list of hypothesized locations. Referring to the example in Table 1, the parsed expressions are underlined. The expression “這裡” (here) will produce a list of all landmarks present in the map in focus, while the expression “這四個大學” (these four universities) will produce a list of all locations of type UNIVERSITY from the map in focus. As regards the pen modality, our framework generates a list of possible semantics based on the gesture type and its (x,y) coordinate(s). Referring again to Table 1, the pointing gesture produces a list of locations whose icons lie in the vicinity (within fifty pixels) of the point’s coordinates, ordered with increasing distances. The circling gesture produces a list of locations whose icons were encircled. As can be seen, our cross-modality integration framework first generates, for each individual modality, partial interpretations represented by a series of listed locations, where each list correspond to an input event (spoken locative reference expression or pen gesture). These are then integrated with a Viterbi alignment algorithm, whose scoring function incorporates semantic compatibility (in terms of numeric and location type features) and temporal order. The integration process is illustrated in Figure 1 and details of the algorithm are described in [5]. Table 2 also presents the unimodal paraphrase based on the multimodal expression in Table 1. Evaluation based on the multimodal test set (342 inputs) shows that the cross-modality integration framework can correctly generate unimodal paraphrases for 97% of the data. The remaining minority with errors is described in [5].

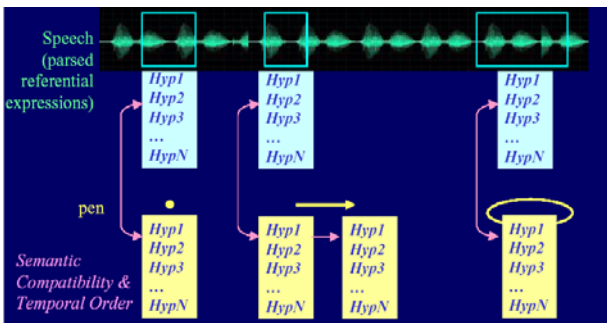


Figure 1. The cross-modality integration framework. Each input event in each modality produces a list of hypothesized locations. These are aligned across modalities while incorporating semantic compatibility and temporal order.

Table 2. Unimodal paraphrase generated by the cross-modality integration framework, based on the multimodal input expression in Table 1.

Unimodal Paraphrase: 我從 <u>身處點</u> 要到 <u>北京航空航大</u> <u>中國地質大學</u> <u>北京科技大學</u> <u>北京醫科大學</u> 一共需要多少時間?
Translation of the spoken utterance: “I want to go from my current location to Beihang University, China University of Geosciences, University of Science and Technology Beijing, Beijing Medical University. How much time will it take?”

### 4. Generated Parallel Multimodal and Unimodal Corpora

We ran the cross-modality integration algorithm on the multimodal user expressions and selected correct unimodal paraphrases (over 97% of the entire data set) to form parallel corpora of multimodal inputs with their semantically equivalent, unimodal counterparts. More specifically, we obtain 725 multimodal and unimodal expression pairs from our training set and 314 pairs from our testing set. Comparative statistics of the multimodal and unimodal inputs are shown in Table 3. 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, 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”, “我目前的地”, 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.

Table 3. Parallel multimodal and unimodal corpora statistics.

	Multimodal Input	Unimodal Paraphrase
Total # of words	9,455	10,286
Average utterance length (in # words)	9.1	9.9
(in # characters)	18	27
Range of utterance lengths (in # words)	1 to 25	1 to 25
(in # characters)	2 to 58	2 to 73
Vocabulary size (# words)	526	545

Table 4. Comparison of the class-trigram perplexities between the parallel multimodal and unimodal test sets.

	Multimodal Inputs	Unimodal Paraphrases
Total # utterances	314	314
# of words	3,157	3,718
<b>Perplexity (PP)</b>	<b>6.03</b>	<b>15.56</b>
# of unigram hits	225 (6.05%)	501 (15.87%)
# of bigram hits	374 (10.06%)	712 (22.55%)
# of trigram hits	3119 (83.89%)	1944 (61.58%)
# of OOVs	30 (0.8%)	46 (1.44%)

#### 4.1. Language Modeling

We pooled the multimodal and unimodal spoken expressions together (1,450 in all) 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 4 other equivalences classes including: ARTICLES, NUMBERS (i.e. numeric expressions), MEASURE\_WORDS and

LOCATION\_TYPE (e.g. the words “*university*”, “*parks*”, etc.) The language model was developed using the CMU SLM toolkit [8]. The resulting model contains 290 unigrams, 1,375 bigrams and 2,795 trigrams. The probabilities are smoothed by Katz backoff smoothing [9] 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 4.

We observe from Table 4 that for the semantically equivalent, parallel multimodal and unimodal corpora, the unimodal paraphrases have significantly higher perplexities. Results from pairwise comparisons between each multimodal (MM) input and its unimodal (UM) paraphrase are in Table 5.

Table 5. Comparison of Per-Utterance Perplexities (PP) between the Multimodal Inputs (MM) and their Unimodal (UM) Paraphrases

	# utterances
$PP_{MM} < PP_{UM}$	264 / 314 inputs (84%)
$PP_{MM} = PP_{UM}$	50 / 314 inputs (16%)
$PP_{MM} > PP_{UM}$	0%

## 5. Data Analysis

These results in Table 5 prompted us to divide the testing data into two subsets, according to  $(PP_{MM}=PP_{UM})$  and  $(PP_{MM}<PP_{UM})$  for further analysis.

Table 6. Illustrative examples from the testing data subset with  $(PP_{MM}=PP_{UM})$ .

<p><b>Example 1:</b> Multimodal Expression, <math>PP_{MM}=3.61</math> (Note redundancy across modalities) S: 從 北郵 到 北航 地質大學 北科大 和 北醫 要多久 P: • • • • • (translation: How much time will it take from <u>BUPT</u> to <u>Beihang</u>, <u>CUG</u>, <u>USTB</u> and <u>BJMU</u>?)</p> <p>Unimodal Paraphrase, <math>PP_{UM}=3.61</math> 從 北京郵電大學 到 北京航空航大 中國地質大學 北京科技大學 和 北京醫科大學 要多久 (translation: How long will it take to go from <u>Beijing Univ. of Post and Telecommunications</u> to <u>Beihang University</u>, <u>China University of Geosciences</u>, <u>University of Science and Technology Beijing</u> and <u>Beijing Medical University</u>?)</p>
<p><b>Example 2:</b> Multimodal Expression, <math>PP_{MM}=4.93</math> (Note ellipsis) S: 最快的交通路線 P: → → → → (translation: The fastest route.)</p> <p>Unimodal Paraphrase, <math>PP_{UM}=4.93</math> 最快的交通路線</p>
<p><b>Example 3:</b> Multimodal Expression, <math>PP_{MM}=654.3</math> S: 我的位置 交通路線 P: • → (translation: <u>my current location</u>. Travel route please.)</p> <p>Unimodal Paraphrase, <math>PP_{UM}=654.3</math> 身處點 交通路線</p>

### 5.1 Category $(PP_{MM}=PP_{UM})$ :

For this category, we found that the majority (33/50=66%) of the expressions involve *redundancy* between the speech and pen modalities. As shown in Example 1 of Table 6, each pair of (x,y) coordinates of each pointing gesture in the multimodal

input *matches with* the abbreviation of the location name that was uttered. The unimodal paraphrase incorporates the full name of each location during generation. However, since our class-based language model gives the same probability values to both the abbreviated and full names of the same location, the per-utterance perplexity values are the same.

Example 2 in Table 6 illustrates the use of ellipsis, which occurred for (16/50=32%) of the cases in this data subset. The subject input four pen strokes that connects four locations and simply uttered “the fastest route”. We interpret that the subject wishes to obtain the fastest route that traverses the four indicated locations. However, the speech modality does not mention the locations at all. Hence the cross-modality integration framework cannot capture the ellipsis and generates a unimodal paraphrase that ignores the pen gestures, resulting in an equal perplexity value. This is an artifact because in reality the multimodal expression conveys a greater amount of information when compared to its unimodal paraphrases. Ellipsis should be a case of complementarity across modalities where certain semantic content appears in one modality and is completely omitted from the other modality.

Example 3 illustrates the occurrence of a spoken locative reference expression that is redundant with the pointing gesture, followed by an ellipsis. Again, we observe equal per-utterance perplexities and the explanations are consistent with the two previous examples.

Redundancy between the speech and pen modalities should be very useful in face of imperfect recognition outputs, e.g. in automatic speech recognition and pen gesture recognition. Handling ellipsis merits further investigation for automatic interpretation of multimodal input.

Table 7. Illustrative examples from the testing data subset with  $(PP_{MM}<PP_{UM})$ .

<p><b>Example 4</b> Multimodal Expression, <math>PP_{MM}=4.53</math> (Note complementarity across modalities) S: 我現在在 這裡 我想分別去 這幾所大學 要多久 P: • S: 有哪些交通線路可以選擇 (translation: I am now <u>here</u>. I want to visit <u>these universities</u>. What are the possible travel routes?)</p> <p>Unimodal Paraphrase, <math>PP_{UM}=6.50</math> 我現在在 北京電影學院 我想分別去 北京航空航大 北京科技大學 中國地質大學 北京醫科大學 有哪些交通線路 路可以選擇 (translation: I am now at <u>Beijing Film Academy</u>. I want to visit <u>Beihang University</u>, <u>China University of Geosciences</u>, <u>University of Science and Technology Beijing</u> and <u>Beijing Medical University</u>. What are the possible travel routes?)</p>
<p><b>Example 5</b> Multimodal Expression, <math>PP_{MM}=5.71</math> (Note complementarity across modalities) First rendition: S: 從 這裡 到 這裡 這裡 這裡 還有 這裡 有什麼交通線路 P: • • • • • (translation: what is the travel route from <u>here</u> to <u>here</u>, <u>here</u>, <u>here</u> and <u>here</u>?)</p> <p>Second rendition: Multimodal Expression, <math>PP_{MM}=9.08</math> (Note redundancy in the first reference expression and complementarity in the remaining four expressions) S: 從 北郵 到 這裡 這裡 這裡 還有 這裡 有什麼交通線路 P: • • • • • (translation: what is the travel route from <u>BUPT</u> to <u>here</u>, <u>here</u>, <u>here</u> and <u>here</u>?)</p>

Unimodal paraphrase  $PP_{UM}=9.21$

從 北京郵電大學 到 北京航空航天大學 北京科技大學  
中國地質大學 還有 北京醫科大學 有什麼交通路線

(translation: what is the travel route from Beijing University of Post and Telecommunications to Beihang University, University of Science and Technology Beijing, China University of Geosciences and Beijing Medical University?)

### 5.2 Category ( $PP_{MM}<PP_{UM}$ ):

The testing data subset with this inequality contains 264 (84%) expressions. We present illustrative examples in Table 7. As shown in Example 4, the speech and pen modalities *complement* each other in specifying a group of intended locations. Either modality alone is semantically ambiguous, e.g. the spoken expression “*here*” that corresponds to the point, or the expression “*these universities*” that correspond to the circle. However, when the semantics across modalities are combined, the semantic meaning 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, the perplexity of the spoken component in the multimodal expression is less than that of the unimodal paraphrase.

Example 5 in Table 7 illustrates the possibility that a multimodal expression can exhibit both redundancy and complementarity in sequential locative reference expressions. The first rendition shows five reference expressions, all of which exhibit complementarity between the speech and pen modalities. There are 242 (92%) similar cases (i.e. complementarity across modalities) in this data subset. The second rendition shows redundancy in the first reference expression, while the remaining four expressions exhibiting complementarity. Hence the per-utterance perplexity rose slightly (c.f. the first rendition) even though both renditions are semantically equivalent. There are 22 (8%) similar cases (i.e. combined redundancy and complementarity) in this data subset. The third rendition is the unimodal paraphrase, which has the highest per-utterance perplexity value.

### 5.3 Findings and Implications

Categorization of the test set based on perplexity values, followed by analysis of the categories enables us to visualize the effects of complementarity and redundancy [2] across the speech and pen modalities in multimodal user inputs.

Complementarity offers expressive power, because the user is free to distribute various parts of the message to different modalities to ease (complex) communication and to reduce cognitive loading [3]. Semantic decoding of an individual modality generates a partial interpretation of the intended message and these partial semantics need to be integrated in order to gain a complete understanding of the user’s intent. This motivates the use of the late semantic fusion architecture for multimodal input interpretation.

Redundancy occurs when both the speech and pen modalities carry the same semantic content. As a preliminary step, the current work only deals with perfect transcriptions of the speech recordings and filtered pen gesture recognition outputs. However, we may conceive that in real applications, the recognition outputs corresponding to different input modalities may be erroneous. Redundancy across modalities motivates the use of mutual disambiguation techniques [10].

In addition, we also observe occurrences of ellipses, where some locative references are omitted from the speech component in the multimodal expression and is expressed only with the pen component. Ellipses motivate further investigations in the syntax of the multimodal language, as well as the use of such multimodal integration approaches as finite-state transducers [11].

## 6. Conclusions

This paper presents a comparative analysis of multi-modal user inputs with speech and pen gestures, together with their semantically equivalent uni-modal (speech only) counterparts. These are generated by a cross-modality framework that applies the Viterbi algorithm to align the speech and pen components in a multimodal expression in order to generate a unimodal paraphrase. We trained a class trigram language model with 1,450 multimodal/unimodal speech utterances and compared the perplexities (PP) between parallel multimodal (MM) and unimodal (UM) test sets (with 314 utterances each). We observe that the speech components of multimodal expressions are generally shorter with lower lexical variability than their unimodal counterparts. Comparison with per-utterance perplexities affirms the relationships of complementarity and redundancy across the speech and pen modalities. One subset of our data exhibits the equality of ( $PP_{MM}=PP_{UM}$ ) and consists mainly of multimodal expressions where speech and pen modalities carry redundant semantics. The other subset exhibits the inequality of ( $PP_{MM}<PP_{UM}$ ) where the speech and pen modalities carry complementary semantics. We also observe the occurrences of ellipsis, where certain semantics appear in one modality but not the other, and forms a special case of complementarity. These observations have implications on the choice of fusion architectures for multimodal input interpretation. Future work will include processing erroneous recognitions and implementation of multimodal fusion.

## 7. Acknowledgements

This work is partially supported by a grant from the HK SAR Government Research Grants Council (Project No. 415105) and is affiliated with the Microsoft-CUHK Joint Laboratory for Human-centric Computing and Interface Technologies.

## 8. References

- [1] Oviatt, S., et al., “Designing the User Interface for Multimodal Speech and Pen-based Gesture Applications: State-of-the-Art Systems and Future Research Directions,” *Human-Computer Interaction*, 15(4), 2000.
- [2] Coutaz, J., et al., “Four Easy Pieces for Assessing the Usability of Multimodal Interaction: The CARE Properties”. *Proceedings of INTERACT*, 1995.
- [3] Oviatt, S., et al., “When Do We Interact Multi-modally? Cognitive Load and Multi-modal Communication Patterns,” *Proceedings of ICMI*, 2004.
- [4] Gupta, A. and T. Anastasakos, “Integration Patterns during Multimodal Interaction,” *Proceedings of Interspeech*, 2004.
- [5] Hui, P.Y. and H. Meng “Joint Interpretation of Input Speech and Pen Gestures for Multimodal Human Computer Interaction,” *Proceedings of Interspeech*, 2006.
- [6] Watanabe, Y., et al., “Semi-synchronous Speech and Pen Input,” *Proceedings of ICASSP*, 2007.
- [7] Oviatt, S., et al., “Integration and Synchronization of Input Modes during Multimodal Human-Computer Interaction,” *Proceedings of CHI*, 1997.
- [8] Clarkson, P. and R. Rosenfeld, “Statistical Language Modeling Using the CMU-Cambridge Toolkit,” *Proceedings of Eurospeech*, 1997.
- [9] Goodman, J. T., “A Bit of Progress in Language Modeling: Extended Version,” *Technical Report MSR-TR-2001-72*, Microsoft Research, Seattle, 2001.
- [10] Oviatt, S., “Multimodal System Processing in Mobile Environments,” *Proceedings of UIST*, New York, 2000.
- [11] Johnston, M. and S. Bangalore, “Finite-state Multimodal Parsing and Understanding,” *Proceedings of the International Conference on Computational Linguistics*, Saarbrücken, 2000.