Multimodal Redundancy Across Handwriting and Speech During Computer Mediated Human-Human Interactions

Edward C. Kaiser, Paulo Barthelmess, Candice Erdmann, Phil Cohen
Adapx
821 Second Avenue, Suite 1150, Seattle, WA, 98104
(ed.kaiser; paulo.barthelmess; candice.erdmann; phil.cohen)@adapx.com
+011 206 428 0800

ABSTRACT

Lecturers, presenters and meeting participants often say what they publicly handwrite. In this paper, we report on three empirical explorations of such multimodal redundancy — during whiteboard presentations, during a spontaneous brainstorming meeting, and during the informal annotation and discussion of photographs. We show that redundantly presented words, compared to other words used during a presentation or meeting, tend to be topic specific and thus are likely to be out-of-vocabulary. We also show that they have significantly higher *tf-idf* (*term frequency–inverse document frequency)* weights than other words, which we argue supports the hypothesis that they are dialogue-critical words. We frame the import of these empirical findings by describing SHACER, our recently introduced Speech and HAndwriting reCognizER, which can combine information from instances of redundant handwriting and speech to dynamically learn new vocabulary.

Author Keywords

Multimodal, Speech, Handwriting.

ACM Classification Keywords

H.5.2 [User Interfaces]: Natural language; Input devices and strategies. I.2.6 [Learning]: Language Acquisition..

INTRODUCTION

Multimodal redundancy occurs when the information in one input mode is semantically the same as information in another input mode, as for example, when a presenter handwrites a phrase like, “Propose your solution,” while also saying it as shown in Figure 1.

A Working Hypothesis of Multimodal Redundancy

In multi-party interactions humans use multiple modes of communication in predictable ways. *Grounding*, for example, is the process by which we attach meaning to symbols we create [15], and *lexical entrainment* [8] is the process of collaboratively adopting dialogue-critical terms for discussing shared referents. In Figure 2, after a meeting facilitator has spoken the phrase, “Information Questions,” while handwriting its abbreviation, *Information Q’s*, on a flipchart, he then pauses, points at the abbreviation and says “right?” These actions ground and entrain the meaning of the handwritten abbreviation.

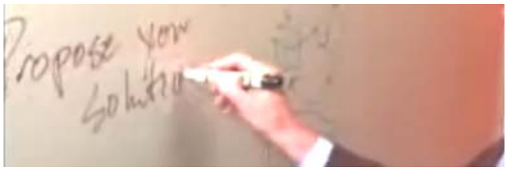


Figure 1: Multimodal Redundancy across handwriting and speech: a whiteboard presenter handwriting *Propose your solution* while also saying, “… *Propose your solution.*”

Herbert Clark’s *Principle of Least Collaborative Effort* [10] argues that humans expend all and only the necessary conversational energy to accomplish dialogue *grounding* and *entrainment* [9, 13]. It is clear that multimodal redundancy — e.g., both handwriting and speaking a term — requires more energy than unimodal communication alone. Therefore, there must be important communicative purposes driving its use.

Our working hypothesis is that people use redundancy as a conversational strategy to bolster their communicative effectiveness by drawing attention to the meanings of dialogue-critical terms. In support of this hypothesis, we consider two derived claims. First, if multimodal redundancy is a general conversational strategy then it should be typical of human-human interaction settings

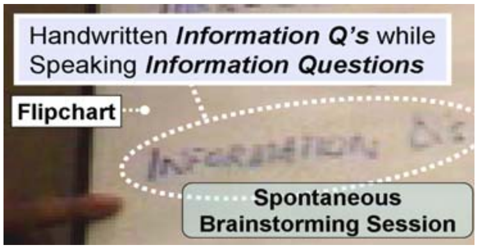

Handwritten *Information Q’s* while
Speaking *Information Questions*
Flipchart
INFORMATION Q's
Spontaneous
Brainstorming Session

Figure 2: During a flipchart brainstorming session the facilitator handwrites *Information Q’s* while also saying, *“Information Questions,”* grounding the abbreviation’s meaning.

CHI 2007, April 28–May 3, 2007, San Jose, California, USA.

where multiple modes can be perceived. Second, if redundantly presented terms are dialogue-critical then they should be measurably more important than non-redundantly presented words.

The Structure of this Paper

In support of this hypothesis we will show that multimodal redundancy is indeed typical in certain communicative situations. Secondly, we will show that words presented redundantly are dialogue-critical, as measured by their greater *tf-idf (term frequency–inverse document frequency)* weights. To introduce these empirical findings we will first discuss related and motivating work. After analyzing multimodal redundancy, we will outline how SHACER, our Speech and HAndwriting recognizer, can leverage its occurrence to dynamically learn important new words, like proper names and their handwritten abbreviations.

RELATED AND MOTIVATING WORK

Multimodal Complementarity versus Redundancy

In multimodal command systems, redundancy has been shown to occur for only between 1%-5% of interactions [14, 29]. Thus the prevailing view in the literature is that for most multimodal commands, *complementarity* rather than *redundancy* is the major organizational theme [28]. In contrast to this prevailing view, Anderson *et al.* [2, 3] have recently found that during computer-mediated, distance-learning lectures, 100% of the presenter's handwriting was accompanied by semantically redundant speech.

This paper confirms and expands upon the findings of Anderson *et al.* We examine three empirical data collections: (1) online whiteboard presentations, (2) a ninety minute spontaneous brainstorming session, and (3) multi-party discussions of photos printed on digital paper. All three of these interaction contexts, as is also true for the studies of Anderson *et al.*, are of human-human interactions where participants share a public writing space.

Multimodal Understanding of Human-Human Interaction

Recently we have introduced a new class of multimodal system. Instead of supporting a direct human-computer interface for command/display turn sequences, it accumulates ambient perceptual observations during structured multi-party interactions, like the construction of a Gantt schedule chart during a meeting [16]. Within this *Ambient-Cumulative-Interface (ACI)* there is no direct human-computer command interface; instead, there is ongoing background computer perception and processing of natural human-human interactions.

For the *ACI* we have implemented for testing SHACER, the perceived interactions occur in public spaces that are shared by the participants — e.g., (a) a shared interactive whiteboard or a piece of digital paper for public sketching and handwriting [5], (b) a shared conversational space for speech captured by close-talking microphones [18]. Participants in this shared public space can be co-located or remotely distributed. The system's function is to unobtrusively collect, recognize, integrate and understand the information it observes in those public spaces, and produce useful background artifacts. For example, our *ACI* Charter application [20] automatically populates an MS Project™ Chart by observing a scheduling meeting and integrating recognized whiteboard Gantt chart sketch and handwriting elements with their associated speech events [16]. Since an *ACI* perceives and processes natural interactions, new terms (i.e., out-of-vocabulary words) will inevitably occur, which are not covered by the system's dictionaries and language models as discussed next.

Out-Of-Vocabulary Words in Natural Speech Contexts

New language constantly emerges from complex, collaborative human-human interactions like meetings, lectures or presentations — such as when a presenter handwrites a new term on a flipchart, like the *Information Q's* abbreviation shown in Figure 2. Fixed vocabulary recognizers tend to fail on such new terms; therefore, we argue that multimodal *ACI* systems need to be able to adapt dynamically to newly introduced vocabulary.

In a recent analysis of lecture speech [12], Glass pointed out that the ideal vocabulary for speech recognition is not the largest vocabulary, but rather one that is both relatively small and has a small Out-Of-Vocabulary (OOV) rate. A small vocabulary minimizes substitution errors, and a small OOV rate minimizes insertion errors. The problem is that in general the size of vocabulary and the rate of OOV are inversely proportional [36]. To illustrate this difficulty Glass compiled a small, 1.5K vocabulary of words common to college lectures in three different course areas, and found that still the 10 most common subject-specific words in each lecture area were OOV. Thus the presence of technical, subject-specific OOV terms makes deriving a vocabulary and language model for lecture speech and other natural speech contexts like meetings a significant challenge.

When a lecture's topic area is known ahead of time, automatic vocabulary expansion can be used [27, 37] to leverage textbooks or targeted web searches to augment recognition dictionaries and language model statistics. Kurihara *et al.*, in their work on the use of *predictive handwriting* during lectures given in Japanese [22], assure full coverage with such methods.

Work in the area of Spoken Document Retrieval (SDR) [11] also must deal with OOV terms. SDR researchers aim to retrieve specific recordings from audio databases, employing queries like those used in searching the web for text-based documents. Saraclar and Sproat [31], while performing speech recognition in support of SDR on a database of six teleconferences with a vocabulary and language model from the Switchboard corpus[1], reported a

[1] Switchboard is an audio recording corpus of spontaneous, two-party telephone conversations on 50 different topics.

12% OOV rate. Using the same vocabulary and language model on Switchboard data itself had only a 6% OOV rate. As the OOV rate increased in moving from Switchboard to Teleconference data so too did the recognition word-error-rate with an attendant loss in precision-recall of spoken document retrieval.

To lessen the harmful effect of OOV terms in SDR, practitioners use sub-word units — like phones or syllables — as a basis for both recognition and query representation [26, 31, 38]. The OOV rate for query words in [23], even with small OOV rates for the SDR data itself, was still found to be 12%. When sub-word units are used as a basis for recognition, the OOV problem is mitigated. Query terms can automatically be transformed into appropriate sub-word units, and sequences of sub-word units can then replace words as the basis for querying an index of spoken documents. SHACER, which is at the core of Charter, our *ACI* multimodal system for processing multi-party Gantt chart meetings, also employs a sub-word unit based recognition and alignment strategy as a basis for learning new vocabulary dynamically.

Multimodality in Learning and Teaching

Moreno and Mayer's theory of multimedia learning [25] is founded on three working assumptions drawn from cognitive psychology [35]: (1) humans have separate processing systems for visual/pictorial versus auditory/verbal channels of information (*dual-channel assumption*), (2) each processing channel has limited capacity (*limited-capacity assumption*), and (3) that meaningful learning requires mental processing in both verbal and visual channels, building connections between them.

Given these assumptions, Mayer and Moreno [24] can explain why presenting text that is also spoken helps students learn more effectively, while presenting visual animations or graphics along with visual and spoken text hurts learning. When the redundancy is across two channels (visual and auditory) then processing proceeds in parallel in both channels and the effect is complementary. When the redundancy is in the same channel (e.g. a visual graphic with accompanying visual text) then the focus of attention must be split overloading cognitive processing and resulting in degraded learning performance.

The import of Mayer and Moreno's findings is that students have better recall and learn more effectively when textual information is presented redundantly in both visual and auditory modes. Next we will show that in some human-human interactions speakers typically present information in just this way, redundantly across both visual and auditory channels, by handwriting words and also saying them.

STUDY OF MULTIMODAL REDUNDANCY

Methodology and Hypothesis

We collected data in three settings: (1) online whiteboard presentations (WP), (2) a spontaneous brainstorming (SB) session, and (3) photo annotation (PA) discussions. The methodology was to annotate all handwriting and speech. For redundancy analysis, the frequency with which handwritten words were accompanied by redundant speech was examined. For *tf-idf* analysis documents were constructed by concatenating the transcripts of both the spoken and handwritten words for a discourse segment.

Term Frequency — Inverse Document Frequency

Tf-idf word weights are commonly used in search and retrieval tasks to determine how important a word is relative to a document [4]. Words that occur with high-frequency in a document, but are relatively rare across the set of documents under consideration, provide a good indication of the document's content [30]. The handwritten abbreviations shown in Figure 3 (e.g., *J*, *LB*) exemplify the relation between dialogue-critical words and *tf-idf* weight. They are dialogue-critical words because without knowing how they are grounded in speech, as shown by the call-outs in Figure 3 (*J* = *Java tier*, *LB* = *Load Balancer*), the underlying visual representation lacks meaning. They also have high *tf-idf* weights because they occur frequently within the presentation, but not so frequently across the entire set of presentations. Thus the abbreviations in Figure 3 are both dialogue-critical and highly weighted.

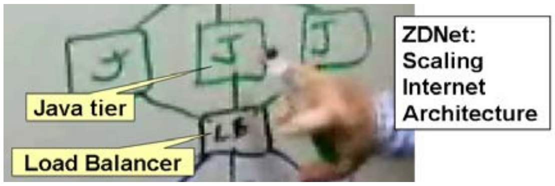


Figure 3: Dialogue-critical words are those whose grounding must be known in order to understand the presentation or discussion (e.g., *J* = "Java tier", *LB* = "Load Balancer").

Corpora Description

Online Whiteboard Presentations (WP)

We examined 34 short (3-4 minutes) whiteboard presentations offered on ZDNet's *At The Whiteboard* site [39]. Figure 1 shows a partial frame from one of these presentations. These presentations discuss various technical and business topics (e.g. Table 4). There was an average of 11.6 handwriting events per presentation, and within those events were 15.9 annotatable handwritten words. In the 34 presentations there were 33 different presenters. The presentation videos were professionally made, and the speakers were in general practiced at presenting information via a whiteboard. Half of the presenters were associated with ZDNet, and half were executives from other companies (e.g. Dell, Intel, etc.). Twenty nine of the presenters were male, and four were female.

Audio and video annotations were done by hand using WaveSurfer's [34] video transcription plug-in. Handwriting was annotated by scrolling the video frame-by-frame to mark the moment of initial pen-down and final pen-up for each handwriting instance. If only one of the pen-up/pen-down events could be clearly seen then the annotator made a best estimate for the other if possible, and if not possible or if neither event could be clearly seen then the handwriting instance was not counted.

Second Scoring: Handwriting Annotation Reliability

A second annotator scored five randomly selected presentations from among the thirty-four, i.e., a 15% random sample. Compared to the first annotator there was a 100% match on what the handwriting events were, a 96% match on the handwritten words within each event, and a 99% match on the spelling of matched words. Between annotators the word start times varied on average by 71 milliseconds and the end times by 49 milliseconds. Rounding up, the handwriting annotation timing accuracy was reliable to within 0.1 seconds.

Spontaneous Brainstorming Session (SB)

Multimodal redundancy also occurs in less formal situations. For example, we recorded a spontaneous brainstorming session, which occurred during a two day planning meeting with 20 participants. Ninety minutes of the session were recorded. Figure 2 is an example of handwriting and speech that occurred during this session. Annotation of handwriting events followed the same procedure used in annotation of the ZDNet whiteboard meetings (see above). For audio transcription, only speech that was associated with a handwriting event was annotated.

All handwriting was performed by the session leader, but the speech associated with the handwriting events was spoken by various participants in the meeting. Only 52% of the speech accompanying the presenter's public handwriting during the brainstorming session was spoken by the handwriter. The other 48% was spoken by seven out of the other 20 meeting participants. The percent of contributions from each of those seven roughly matched their positions in the organizational hierarchy underlying the meeting. So, the project manager's contributions were greatest (14%) followed by those of the project lead (9%), team leads (9%, 5%, 5%) and then of the project engineers (5%, 3%).

Terminology

In a document, each unique word is referred to as a word *type*, while each individual word occurrence is referred to as a word *token*. If while saying "hand over hand" a presenter also wrote the word *hand*, then concatenating the speech and handwriting transcripts would yield the word token list, "hand over hand *hand*," with three tokens of the word type, *hand*. We refer to the word types in this combined token list as *overall* types (i.e., *ALL*) because they can originate from either speech or handwriting. The subset of *ALL* word types that were handwritten are *HW* word types. The subset of *HW* types that were redundantly handwritten and spoken are *RH* types.

In natural language processing tasks, a *stop-list* typically contains closed class words like articles, prepositions, pronouns, etc., which tend to occur with equal relative frequency in most documents. When computing *tf-idf* weights [4], the stop words (i.e., words occurring on the *stop-list*) are removed from consideration, because they tend to add little to the determination of which words are important representatives of a particular document.

Photo Annotation (PA) using Digital Paper and Pen

In [6] we reported on some aspects of a pilot study in which photos printed on digital paper were discussed and simultaneously annotated with a digital pen (Figure 4). There were four annotation sessions. In this paper we further analyze data from the two native English speakers' sessions. All speech for these photo annotation sessions was hand annotated, but the handwriting gestures were automatically captured via digital paper and pen (Figure 4).

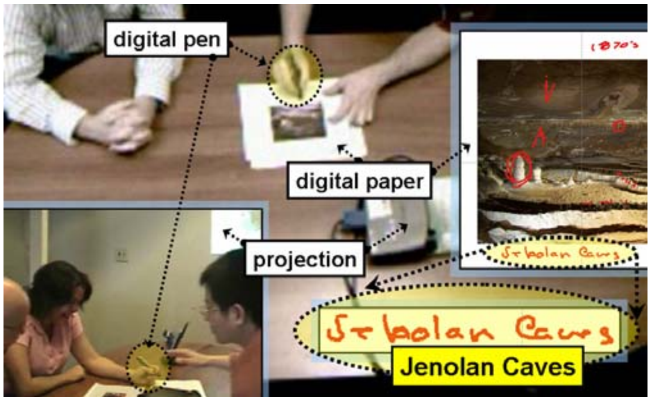


Figure 4: For travelogue photos printed on digital paper the handwriter labels the place name, *Jenolan Caves*, with a digital pen, while also saying, *"... this is the Jenolan Caves."*

Participants were asked to choose some photos they'd like to discuss (nine and ten photos each for the sessions we examine here). They then spoke about their photos to a small group of others (Figure 4), having been told that they could annotate freely and that the software would process their annotations so they would get back labeled photos. Photos were automatically projected on a shared display (Figure 4, lower left inset, *projection* space), since audience members sitting across the table could not easily see the paper versions. The projected images were updated when the digital pen touched a photo sheet [6].

Study Results

Amount of Handwriting

Previously, Kurihara *et al.* found that as much as 18% of lecture time was spent handwriting [22]. For the ZDNet whiteboard presentations examined here, the presenters spoke on average for 192.9 seconds (stddev = 44.3 seconds) and handwrote on average for 38.9 seconds (stddev = 20.9 seconds). Thus, on average 21.3% (stddev = 13.4%) of presentation time was spent in handwriting.

Redundancy

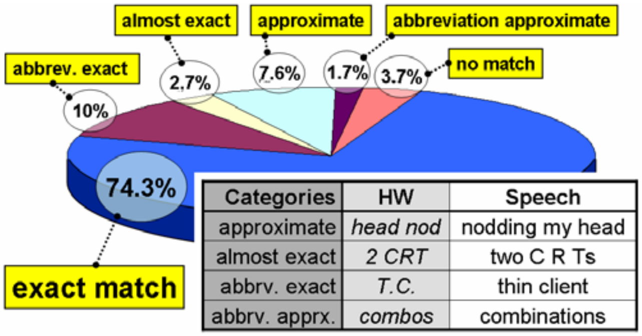
Table 1 shows the number of handwritten words that occurred in each of the three corpora (*HW* row), along with the number of handwritten words that were also spoken redundantly (*RH* row). The bottom row of Table 1 shows the percent of handwritten words that were spoken redundantly (*RH/HW* row). The average number of handwritten words accompanied by redundant speech over all three corpora was 96.5%. These results support the claim, which is derived from our working hypothesis, that multimodal redundancy is typical of human-human interaction settings where multiple modes can be perceived.

	WP	SB	PA	TOTAL
Handwritten Words (**HW**)	492	41	155	**688**
Redundantly spoken **HW** (**RH**)	479	40	145	**664**
Redundancy (**RH/HW**)	97.4%	97.6%	93.5%	**96.5%**

Table 1: Redundancy rates across ZDNet whiteboard presentations (WP), the spontaneous brainstorming (SP) session, and the photo annotation (PA) discussions.

Figure 5 shows the types of redundant matches that occurred, averaged over all three corpora. The preponderance of matches were *exact* lexical matches (74.3%), where the handwritten terms were spoken exactly as written. *Abbreviation exact* matches were defined as standard abbreviations that exactly match their expansions in speech (10% — e.g., Fig. 2, Fig. 5 inset). *Almost exact* matches differ only in number or tense (2.7%). *Approximate* matches differ in word order or form (Fig. 5 inset), or have extra or missing words (7.6%), as is also true for the spoken expansions of *abbreviation approximate* matches (1.7%). Category examples are shown in the *Categories* inset of Figure 5. For the ZDNet corpus by itself the percentage of abbreviations was 44.3%, which was much higher than for the other corpora.

Our result of 74.3% *exact* match with 96.5% overall redundancy closely parallels the 74% *exact* match and 100% redundancy found earlier by Anderson *et al.* [2]. However, Anderson *et al.* examined only 54 instances of handwriting. This paper analyzes an order of magnitude more data — 688 handwriting instances. Our findings are thus numerically more significant. We also examine three different scenarios, none of which was based on the use of a tablet PC as in the study by Anderson *et al.*



Categories	HW	Speech
approximate	*head nod*	nodding my head
almost exact	*2 CRT*	two C R Ts
abbrv. exact	*T.C.*	thin client
abbrv. apprx.	*combos*	combinations

Figure 5: Redundancy category breakdown averaged across ZDNet whiteboard presentations (WP), the spontaneous brainstorming (SP) and photo annotation (PA) sessions.

Of these six different types of redundancy, SHACER can currently take advantage of three — (1) *exact*, (2) *abbreviation exact*, and (3) *almost exact* redundancies. These three categories represent 87% of the handwriting events reported on in this paper. Within the *no match* category (3.7%) there was a sub-category dubbed *semantic matches*. These are cases in which, for example, a narrator while writing the name of a family member (e.g. *Donald*) says both the relationship and name of that family member (e.g., "my son, Donald"), and then later while again writing the name says only the relationship, "my son." Such *semantic matches* occurred in about 1% of redundant instances. Both *semantic matches* and *approximate matches* could conceivably be processed by SHACER in the future.

Redundancy Timing

Understanding the temporal relationship between handwriting events and redundant speech is important. If they likely to be temporally close then the search space for aligning and detecting such redundancies can be reduced.

Sequential (24%)	**Simultaneous (76%)**		
Writing First	**Speech Precedes**	**Writing Precedes**	**Neither Precedes**
HW__ S__ (16%)	**S______ HW___ (1%)**	**S____ HW_____ (6%)**	**S____ HW_____ (0%)**
Speech First	**S______ HW__ (11%)**	**S__ HW______ (39%)**	**S____ HW__ (2%)**
S__ HW___ (8%)	**S_____ HW____ (1%)**	**S_____ HW____ (15%)**	**S___ HW______ (1%)**

Table 2: Temporal categories by precedence (for ZDNet corpus). Note that 24% of instances are sequential (left), with no overlap between handwriting (W) and speech (S).

Following Oviatt *et al.* [29] we have examined the temporal integration patterns of redundantly delivered inputs. For the 34 presentations of the ZDNet corpus, we found that 24% of redundant inputs were presented sequentially with either handwriting occurring first followed by speech (Table 2, *Writing First* – 16%), or speech occurring first (8%). For simultaneous (over-lapping) constructions, which were 76% of instances, speech preceded handwriting in 13% of cases, handwriting preceded speech in 60% of cases, and neither preceded in 3% of cases (timing accurate to 0.1 sec). The tendency of handwriting to precede speech was significant by Wilcoxon signed ranks test, T+=524.5 (N=32), p<0.0001, one-tailed.

When we superimpose the timing data from the spontaneous brainstorming (SB) session onto to that of the ZDNet presentations (Figure 6), the timing contours are closely matched (see *leader* and *ZDNet* lines). Figure 6 shows the number of seconds from start-of-handwriting to

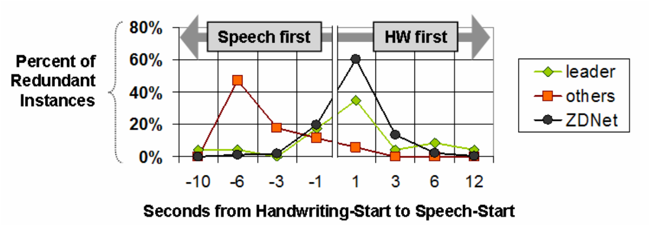


Figure 6: The number of seconds by which the start of handwriting (HW) preceded the start of speech. Negative values mean that speech preceded HW. The plot includes data from both the brainstorming (SP) session's *leader* and *others*.

the start-of-speech. Negative values mean speech occured first. During the spontaneous brainstorming session, when handwriting was spoken redundantly by others rather than by the leader, there was a marked shift in the peak amount of time by which speech preceded handwriting (Figure 6, *others* line). Thus, when speaking about his own handwriting the leader's timing pattern closely matched that of the average ZDNet presenter — with handwriting slightly preceding speech and simultaneously overlapping it. However, when the speech of other meeting participants was reflected in his handwriting, then that handwriting occurred a few seconds after the terms had been spoken.

Each input pair of the sequential inputs shown in Table 2 is by the same ZDNet presenter. Of these inputs 33% were speech followed by handwriting, a pattern which for speech and sketched graphics in Oviatt *et al.* [29] occurred for only 1% of the sequential inputs. This may suggest that because handwriting requires more cognitive effort than sketching it is therefore delayed in presentation compared to simple locative sketches.

For sequential patterns, a preponderance of inter-modal lag times (i.e., the time from the end of first mode to start of next mode) was less then 2 seconds: 80% in the speech first case, and 76% in the handwriting-first case. For the speech-first condition all lags were within 4 seconds. For the handwriting-first condition 8% of the lags were longer than 4 seconds, with the longest being a full minute and a half.

Redundancy and Projected Out-of-Vocabulary Words

As discussed above, Glass *et al.* [12] examined the nature of OOV words in a small general vocabulary common to a training set of lectures. They found that subject-specific words from lectures were not well covered and often missing even from the vocabularies of larger corpora like Broadcast News[2] and Switchboard. Here we perform a similar examination of word type sharing across the 34 presentations of the ZDNet whiteboard presentation corpus. If we choose a vocabulary of all words that are not presentation-specific what level of coverage will there be?

Table 3 shows the results of examining the number of shared word tokens and word types along with the number of shared handwriting (*HW*) types. Row 1 of Table 3 shows that across all 34 presentations 15.81% of word tokens were shared, while only 0.25% of word types and just 1.03% of handwritten types were shared commonly. This illustrates the effect of not removing *stop-list* words (*no SL*): a small percentage of word types (e.g. closed-class words) accounts for a relatively large number of shared tokens. With no stop list removal the average number of handwriting types per presentation was 15.9. There were 209 average overall word types per presentation. The percent of overall word types occurring in only one presentation (Table 3, row 2, *no SL*) was 59.95%, and of handwritten types was 48.32%. Such presentation-specific words will be OOV for a shared common vocabulary.

	Num. Shared Presentations	Tokens Shared	Types Shared	HW Types Shared	Avg. HW
1	34	15.81%	0.25%	1.03%	15.9
2	1 *(no SL)*	12.54%	59.95%	48.32%	15.9
3	1 *(SL)*	27.04%	64.96%	51.66%	14.6
4	1 *(SL + 20k)*	53.46%	88.15%	76.79%	4.4
5	1 *(SL + 170k)*	65.65%	88.29%	82.76%	3.0

Table 3: Percent of word tokens and types common to a Number of Shared Presentations. Percent of handwritten (HW) types commonly shared is also given, as well as the average number of HW types (Avg. HW) per presentation

In the lower three rows of Table 3 (rows 3-5) we show the percentage of shared types remaining after basic stop list removal and with increasingly larger removal dictionaries: *SL* = basic stop list; *20k* = a 20,000 word dictionary of the most common words in a corpus of meetings; and *170k* = a 170,000 word dictionary from the Festival Speech Synthesis Toolkit [7]. As the number of common word types removed increases the remaining word types tend to be more and more presentation-specific. However, it can be seen that as dictionary size increases the number of average handwritten types per presentation (not removed by the dictionary) decreases from 14.6 (row 3) to only 3 (row 5). With a large general dictionary (e.g. 170k) the roughly 7 presentation-specific handwritten types present in row 3 (51.66% * 14.6 ≈ 7) are reduced to just 2 in row 5 (82.76% * 3 ≈ 2). Thus using large dictionaries does reduce the number of presentation-specific words that are likely to be OOV; but, as Glass [12] has pointed out, this is not ideal. Larger dictionaries require more computational resources and are susceptible to higher word-error rates due to substitutions.

Perhaps, if we had many more training presentations to examine we could hope to find a shared vocabulary with fewer OOVs. Figure 7 shows a power regression prediction that addresses this question. As in row 2 of Table 3, the data points in the left side of Figure 7 are computed with no stop list removal. To accumulate these data points we processed every possible subset of our corpora (i.e., 1 meeting out of the 34, 2 meetings out of the 34, 3 meetings out of the 34, etc.), asking for each increasingly larger subset how many

[2] Recorded and transcribed television and radio broadcasts.

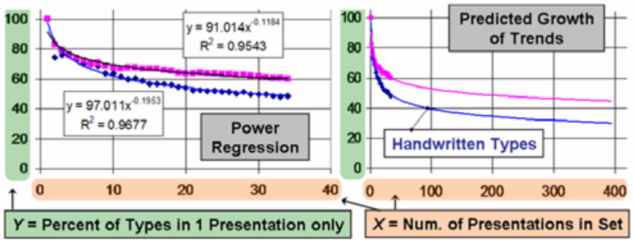


Figure 7: A power regression prediction to examine the percentage of handwritten word types occurring in only one presentation, given an increasing number of presentations. Upper line = overall types, Lower line = handwritten types.

overall and handwritten types occurred in only one presentation. The plot shows that the percent of presentation-specific overall word types (upper trend line) and handwritten types (lower trend line) decreased steadily as set size increased. But the rate of decrease appeared to be leveling off around 40%. The power regressions were computed in MS Excel. The R-squared values indicate goodness of fit: 0.95 for overall and 0.97 for handwritten types. Regression equations are shown in Figure 7.

In the plot on the right side of Figure 7 we have extended the power regression formulas from the left side plot to see what rate of presentation-specific handwritten words might still be present after examining a training set ten times the size of our ZDNet corpus. Trend lines are extended to 360 presentations. Even with this simulated order of magnitude larger training set there was still about 30% of handwritten types predicted to be presentation-specific (Figure 7, right side, lower trend line). Thus for natural speech contexts, even when a large training corpus is available, these findings suggest that as much as a quarter or more of redundant handwritten words would still be presentation-specific and thus out-of-vocabulary. In the next section we will show that such redundant handwritten words, which are likely to be highly presentation-specific, are indeed the dialogue-critical words that one would want to recognize and understand for later retrieval tasks.

Redundancy,TF-IDF Weight and Retrieval Searching

In earlier work [6] we showed that for photo annotation sessions, redundantly introduced words had a 90% higher average frequency than overall word types. In this paper we calculate the average *tf-idf* weights of overall word types (*ALL*) versus redundant handwritten word types (*RH*), for not only the two native English-speakers' photo annotation sessions but also for the ZDNet corpus. For this combined data set, Figure 8 shows the average *tf-idf* weight increase for *RH* types compared to *ALL* types. These strikingly higher *tf-idf* weights for *RH* types — 128% higher with no stop-word removal and 70.5% higher with stop-word removal — were significant by Wilcoxon signed ranks test, T+=561 (N=33), p<0.0001, one-tailed.

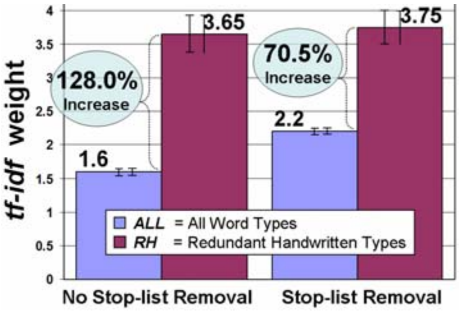


Figure 8: Average *tf-idf* weight increases for redundant handwritten word types (*RH*) versus all word types (*ALL*) for both ZDNet and Photos corpora. Increases are significant.

Table 4 shows examples from three ZDNet presentations of the top ten most highly *tf-idf*-weighted word types (after basic stop list removal). In some presentations – like the left-most, *Detecting Greynets* – all of the top ten are redundantly presented words. Even for those presentations with relatively lower percentages of redundant handwritten

	Detecting Greynets					Rootkits					Network-Centric Computing				
	RH	WGHT	TF	DF	term	RH	WGHT	TF	DF	term	RH	WGHT	TF	DF	term
1	RH	5.92	2	1	adware	RH	13.79	19	1	rootkits	RH	6.69	4	2	client
2	RH	5.92	2	1	block	RH	9.12	5	1	detectors	__	5.92	2	1	environment
3	RH	5.92	2	1	conferencing	__	7.34	3	1	trick	RH	5.92	2	1	mainframe
4	RH	5.92	2	1	enable	RH	5.92	2	1	blacklight	__	5.92	2	1	series
5	RH	5.92	2	1	evasive	RH	5.92	2	1	ghostbuster	RH	5.92	2	1	thin
6	RH	5.92	2	1	hygiene	RH	5.92	2	1	invisible	__	5.03	3	3	computer
7	RH	4.75	2	2	compliance	RH	4.75	2	2	anti	RH	4.05	5	7	server
8	RH	4.75	2	2	deployed	RH	4.75	2	2	spyware	__	3.50	1	1	addresses
9	RH	4.75	2	2	policies	RH	4.75	2	2	virus	__	3.50	1	1	architect
10	RH	4.75	2	2	spyware	__	4.06	2	3	pieces	__	3.50	1	1	attention
In Top 10 TFW	general types		10 / 149 = 6.70%			general types		10 / 131 = 7.60%			general types		10 / 139 = 7.20%		
	RH types		10 / 21 = 47.62%			RH types		8 / 13 = 61.54%			RH types		4 / 5 = 80.0%		

Table 4: Top 10 word types ranked by *tf-idf* weight (WGHT) for three presentations from the ZDNet corpus. Key: TFW = Term Frequency Weightings, RH = Redundantly spoken Handwriting, TF = Term Frequency, DF = Document Frequency.

(*RH*) words in the top ten – as for the right-most, *Network-Centric Computing* – it can be seen that *RH* words as a class are much more likely to be representative terms than *non-RH* words as a class (bottom rows, Table 4, *In Top 10 TFW*). On average for all 34 meetings only 7.66% of overall types are present in the top ten most highly weighted words for a presentation. But of the redundant handwritten (*RH*) types, 61.47% are present in the top 10, which represents 48.64% of all top ten words for all presentations. Thus, the likelihood of a word being one of the top 10 most highly weighted words is less than 1 in 10 (7.66%) for overall word types, while for *RH* word types it is about 5 in 10 (48.64%), meaning that *RH* words as a class are significantly more representative of a presentation than *non-RH* words (by Wilcoxon signed ranks test, T+=593 (N=33), p<0.0001, one-tailed).

Similarly, on average, for all 19 individual photo discussions, just 11.5% of *ALL* types are present in the top 10 most highly weighted words. But of the *RH* types, fully 81.77% were ranked in the top 10, which represents 48.95% of all top ten words for all photo discussions.

Table 4 shown that redundantly handwritten and spoken word types (*RH*) as a class are better representatives of their respective presentations or discussions than other words. Since they have significantly higher *tf-idf* weights than other words, they should be effective search query terms. To test this claim we performed retrieval queries on an indexed directory of speech and handwriting transcript documents, one such document for each presentation in the ZDNet corpus. The search engine we used was a state-of-the-art, open-source search application called *Seekafile* [32], which works on both small and large data sets.

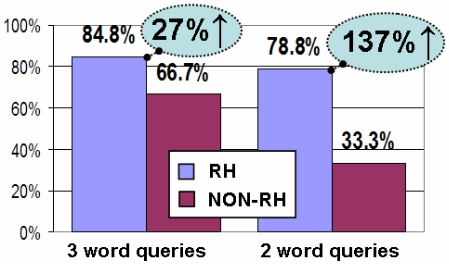


Figure 9: Retrieval accuracy using randomly selected three and two word queries, with words being chosen from the sets of redundantly presented handwritten word types (*RH*) and non-redundantly presented word types (*non-RH*).

We performed searches with both three-word and two-word queries (Figure 9). For each presentation the query words were randomly chosen from either the set of redundantly handwritten and spoken words (*RH* bars in Fig. 9) or from the set of words that were not redundantly presented (*non-RH* bars in Fig. 9). Retrieval accuracy measured how often the best-scoring retrieval result was the correct result.

The outcome for three word queries (Figure 9, left side) shows that words from the *RH* set yielded 84.8% retrieval accuracy while *non-RH* words yielded 66.7% accuracy. Thus for randomly chosen three word queries the retrieval accuracy was 27% higher using *RH* rather than *non-RH* words (marginally significant by Wilcoxon signed ranks test, p<0.0655).

For two word queries the right side bar chart in Figure 9 shows that randomly chosen words from the *RH* set yielded 137% higher accuracy than randomly chosen words from the *non-RH* set. *RH* accuracy was 78.8%, while *non-RH* accuracy was only 33.3%. Thus for two-word queries the retrieval accuracy was significantly higher using *RH* as opposed to *non-RH* words (Wilcoxon signed ranks test, T-=246, N=23, p<0.0001). These results support the claim that redundantly presented words, which as a class have significantly higher *tf-idf* weights than non-redundantly presented words, are more effective search query terms.

Study Implications

From the work of Moreno and Mayer [25] on multimedia learning we know that redundantly presented words are easier to recall, and support better learning. This means that, after seeing redundantly presented words during a presentation, those words will later come to mind more readily for use in retrieval queries. We have also shown that redundant words are likely to be presentation-specific and thus OOV. Allauzen & Gauvain in [1] have reported that up to 70% of OOV words are named entities, like proper names. In the section below we show how SHACER can leverage multimodal redundancy to learn OOV proper names and their handwritten abbreviations. Understanding these redundant OOV terms is critical for background understanding of a Gantt chart created during a meeting.

SHACER

SHACER's goal is to dynamically learn OOV terms (including their handwritten abbreviations), as they are presented redundantly during the course of an interaction. In the lower pane of Figure 10, the *recognized inputs* row shows recognition results from both the handwriting recognizer and speech recognizer. For example, the *Fred Green* handwriting, which labels the Gantt chart taskline beneath it, is incorrectly recognized as *i-redesign* (due to an ink-skip), and the speech is incorrectly recognized as, “Fred's Green,” because the proper name is not in the system's language model. After SHACER combines the redundant handwriting and speech information, both labels were corrected as shown in the *integrated inputs* row. In [19] SHACER corrected 22 of 29 such Gantt chart labeling errors across its development test set — a significant 76% relative error rate reduction (McNemar test, p<=2.98e-06).

In the upper pane of Figure 10, the *recognized inputs* row shows two labels beneath a diamond-shaped Gantt chart milestone. Neither of these handwritten abbreviations (*CB* and *Fig*) is semantically grounded. They have no call-outs that indicate their spoken expansions, and are thus considered incorrect abbreviation recognitions. Combining

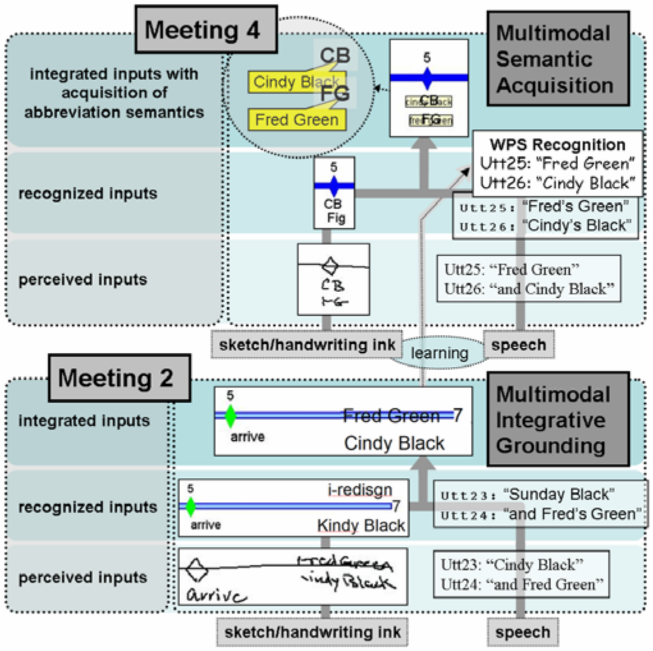


Figure 10: SHACER example: learning abbreviation expansions through *Multimodal Integrative Grounding* (Meeting 2, lower pane) followed by *Multimodal Semantic Acquisition* (Meeting 4, upper pane). WPS = Word/Phrase-Spotter recognition of new terms (e.g. *Fred Green* and *Cindy Black*) previously enrolled during Meeting 2 (lower pane).

redundant information not only corrected the letter string interpretation of *Fig* to *FG* but also grounded both abbreviations to their spoken meanings (*FG* = *Fred Green*, *CB* = *Cindy Black*) [18-21] (Figure 10, upper pane, *integrated inputs* row). On a held-out set of five related test meetings SHACER corrected 6 of 16 such abbreviation label errors, for a significant 37% absolute reduction of error rate (McNemar test, p<=0.03). These results clearly support our earlier findings in [17] that combining information from redundant handwriting and speech is significantly more reliable for the recognition of Gantt chart labels than depending on either mode alone.

SHACER uses sub-word unit recognition for characterizing OOV terms, similar to recognition techniques used in spoken document and spoken name retrieval systems [31, 33]. However, SHACER's aim is not retrieval but rather dynamic learning, which means recognizing the spelling, pronunciation and local semantics of new terms, enrolling them into dictionaries and language models as the system is running, and thus improving the system's accuracy and understanding over time and usage.

Integration and Learning from Multimodal Redundancy

Figure 10 shows a sequence of two meetings during which SHACER learns the expanded meaning of two new OOV terms and their abbreviations. The lower pane of Figure 10 (*Meeting 2*) illustrates *Multimodal Integrative Grounding*, in which the spelling and pronunciation of new terms are dynamically learned by integrating redundant information from handwriting and speech. When a new term (e.g. *Fred Green*) has been dynamically learned, then its spelling and pronunciation are enrolled into a special Word/Phrase-Spotting (WPS) Recognizer. Information stored in that WPS recognizer can be serialized and thus carried across meeting boundaries. When an enrolled new term is spoken later, as for example in *Meeting 4* (Figure 10, upper pane), it is recognized by the WPS recognizer, and its spelling is compared to temporally nearby handwriting. For the cases shown in Figure 10 the nearby handwritten labels are first-letter abbreviations, *FG* and *CB*. The WPS spoken recognition, together with the nearby first-letter abbreviation matches, triggers the association to spoken semantics [18]. We call this associative process *Multimodal Semantic Acquisition*, because learned semantics carried in one mode — WPS speech recognition of *Fred Green* or *Cindy Black* — are dynamically acquired by new symbols in another mode (Fig. 10, upper, *Integrated inputs with acquisition of abbreviation semantics* row).

SHACER learns from as little as a single instance of multimodal redundancy, but it can also benefit from repeated associations. Currently such repetitions expand the list of pronunciation variations enrolled in SHACER's Word/Phrase-Spotting recognizer, thus improving the chances of subsequent recognitions.

Boot-Strapped Learning

Multimodal redundancy integration, in the two phases of *Multimodal Integrative Grounding* and *Multimodal Semantic Acquisition* (Fig. 10), supports boot-strapped learning. The system learns dialogue-critical OOV proper names and abbreviations on its own, with no supervision but that provided by multimodal redundancy itself.

CONCLUSION

Our working hypothesis was that people used multimodal redundancy to focus attention on important words. Derived from that hypothesis was the claim that if multimodal redundancy is a general communicative strategy, then it should be typical in human-human interaction settings. Averaged across three separate contexts we found that 96.5% of handwritten words were also spoken redundantly, which supports the view that such redundancy is typical.

Furthermore we have shown that (1) as much as a quarter of redundantly presented handwritten words are likely to be out-of-vocabulary in relation to ideally sized recognition vocabularies, regardless of training set size, (2) that such redundancies are good mnemonic representatives of a presentation (based on findings from the multimedia learning community), and (3) that as a class they are significantly more representative of a presentation than other non-redundant word types, as measured by higher *tf-idf* weights and significantly better accuracy in search retrieval results. The second claim derived from our working hypothesis was that if redundant words are dialogue-critical they should be measurably more important than other words. These results support this claim.

In describing our work with SHACER, we have shown that redundantly presented terms are dynamically learnable by unsupervised, boot-strapped methods. Such terms are thus at once likely to be OOV and also likely to be dynamically learnable. We believe that dynamic learning of redundantly presented terms is a viable and important way forward towards more adaptive multimodal interfaces.

ACKNOWLEDGEMENTS

This material is based upon work supported by the Defense Advanced Research Projects Agency (DARPA) under Contract No. NBCHD030010. Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the DARPA or the Department of Interior-National Business Center (DOI-NBC).

REFERENCES

1. Allauzen, A. and J.-L. Gauvain. Open Vocabulary Asr for Audiovisual Document Indexation. *ICASSP 2005.*
2. Anderson, R., C. Hoyer, C. Prince, J. Su, F. Videon, and S. Wolfman. Speech, Ink and Slides: The Interaction of Content Channels. *ACM Multimedia*, (2004).
3. Anderson, R.J., R. Anderson, C. Hoyer, and S.A. Wolfman. A Study of Digital Ink in Lecture Presentation. *CHI '04*, (2004).
4. Baeza-Yates, R. and B. Ribeiro-Neto, *Modern Information Retrieval*: Addison-Wesley, 1999.
5. Barthelmess, P., E.C. Kaiser, X. Huang, and D. Demirdjian. Distributed Pointing for Multimodal Collaboration over Sketched Diagrams. *ICMI 2005.*
6. Barthelmess, P., E.C. Kaiser, X. Huang, D. McGee, and P. Cohen. Collaborative Multimodal Photo Annotation over Digital Paper. *ICMI'06,* (2006).
7. Black, A., P. Taylor, and R. Caley, The Festival Speech Synthesis System: System Documentation, *Technical Report HCRC/TR-83*. 1998, Human Communication Research Centre.
8. Brennan, S. Lexical Entrainment in Spontaneous Dialogue. *International Symposium on Spoken Dialogue*, (1996), 41-44.
9. Chai, J.Y., Z. Prasov, J. Blaim, and R. Jin. Linguistic Theories in Efficient Multimodal Reference Resolution: An Empirical Investigation. *IUI '05,* (2005), 43-50.
10. Clark, H.H., *Using Language*: Cambridge University Press, 1996.
11. Garofolo, J., G. Auzanne, and E. Voorhees. The Trec Spoken Document Retrieval Track: A Success Story. *RAIO-2000: Content-Based Multimedia Information Access Conference*, (2000), 1-20.
12. Glass, J., T.J. Hazen, L. Hetherington, and C. Wang. Analysis and Processing of Lecture Audio Data: Preliminary Investigations. *HLT-NAACL Workshop on Interdisciplinary Approaches to Speech Indexing and Retrieval*, (2004).
13. Grice, H.P., *Logic and Conversation*, in *Speech Acts*, P. Cole and J. Morgan, Eds., Academic Press: 1975, New York. 41-58.
14. Gupta, A.K. and T. Anastasakos. Dynamic Time Windows for Multimodal Input Fusion. *INTERSPEECH-2004*, (2004), 1009-1012.
15. Harnad, S., The Symbol Grounding Problem. *Physica D* **42**, (1990), 335-346.
16. Kaiser, E., D. Demirdjian, A. Gruenstein, X. Li, J. Niekrasz, M. Wesson, and S. Kumar. Demo: A Multimodal Learning Interface for Sketch, Speak and Point Creation of a Schedule Chart. *ICMI '04*, (2004). Kaiser, E.C. Multimodal New Vocabulary Recognition through Speech and Handwriting in a Whiteboard Scheduling Application. *IUI '05*, (2005), 51-58.
17. Kaiser, E.C. Shacer: A Speech and Handwriting Recognizer. *ICMI, Multimodal, Multiparty Meeting Processing*, (2005).
18. Kaiser, E.C. Using Redundant Speech and Handwriting for Learning New Vocabulary and Understanding Abbreviations. *ICMI '06*, (2006), 347-356.
19. Kaiser, E.C. and P. Barthelmess. Edge-Splitting in a Cumulative Multimodal System, for a No-Wait Temporal Threshold on Information Fusion, Combined with an under-Specified Display. *INTERSPEECH 2006.* Kaiser, E.C., P. Barthelmess, and A. Arthur. Multimodal Play Back of Collaborative Multiparty Corpora. *ICMI, Multimodal, Multiparty Meeting Processing Workshop*, (2005).
20. Kurihara, K., M. Goto, J. Ogata, and T. Igarashi. Speech Pen: Predictive Handwriting Based on Ambient Multimodal Recognition. *CHI '06,* (2006), 851 - 860.
21. Logan, B., P. Moreno, J.-M.V. Thong, and E. Whittaker. An Experimental Study of an Audio Indexing System for the Web. *ICSLP*, (2000).
22. Mayer, R.E. and R. Moreno, Nine Ways to Reduce Cognitive Load in Multimedia Learning. *Educational Psychologist* **38**, 1, (2003), 43-52.
23. Moreno, R. and R.E. Mayer, Verbal Redundancy in Multimedia Learning: When Reading Helps Listening. *Jour. of Educational Psychology* **94**, 1, (2002), 156-163.
24. Ng, K. and V. Zue, Subword-Based Approaches for Spoken Document Retrieval. *Speech Communication* **32**, 3, (2000), 157-186.
25. Ohtsuki, K., N. Hiroshima, M. Oku, and A. Imamura. Unsupervised Vocabulary Expansion for Automatic Transcription of Broadcast News. *ICASSP '05*, (2005).
26. Oviatt, S., Ten Myths of Multimodal Interaction. *Communications of the ACM* **42**, 11, (1999), 74-81.
27. Oviatt, S.L., A. DeAngeli, and K. Kuhn. Integration and Synchronization of Input Modes During Multimodal Human-Computer Interaction. *CHI '97,* (1997).
28. Salton, G. and C. Buckley, Term-Weighting Approaches in Automatic Text Retrieval. *Information Processing & Management* **24**, 5, (1988), 513-523.
29. Saraclar, M. and R. Sproat. Lattice-Based Search for Spoken Utterance Retrieval. *HLT/NAACL*, (2004), 129-136.
30. Seekafile, Http://Www.Seekafile.Org/.
31. Sethy, A., S. Narayanan, and S. Parthasarthy. A Syllable Based Approach for Improved Recognition of Spoken Names. *ISCA Pronunciation Modeling*, (2002).
32. WaveSurfer, Http://Www.Speech.Kth.Se/Wavesurfer/, Department of Speech, Music and Hearing, Royal Institute of Technology (KTH).
33. Wickens, C.C., Multiple Resources and Performance Prediction. *Theoretical Issues in Ergonomics Science* **3**, 2, (2002), 159-177.
34. Woodland, P.C., S.E. Johnson, P. Jourlin, and K.S. Jones. Effects of out of Vocabulary Words in Spoken Document Retrieval. *ACM Conference on Research and Development in Information Retrieval*, (2000), 372-374.
35. Yu, H., T. Tomokiyo, Z. Wang, and A. Waibel. New Developments in Automatic Meeting Transcription. *ICSLP*, (2000).
36. Yu, P., K. Chen, C. Ma, and F. Seide, Vocabulary-Independent Indexing of Spontaneous Speech. *IEEE Transactions on Speech and Audio Processing* **13**, 5, (2005), 635- 643.
37. ZDNet, Zdnet Whiteboard Videos, in *http://news.zdnet.com/2036-2_22-6035716.html*.