Summarization of Spoken Language — Challenges, Methods, and Prospects

Klaus Zechner
Language Technologies Institute
Carnegie Mellon University
zechner@cs.cmu.edu

January 14, 2002

1 Introduction

While the field of summarizing written texts has been explored for many decades, gaining significantly increased attention in the last five to ten years, summarization of spoken language is a comparatively recent research area. As the amount of spoken audio databases is growing rapidly, however, we predict that the need for high quality summarization of information contained in this medium will rise substantially. Summarization of spoken language may also aid the archiving, indexing, and retrieval of various records of oral communication, such as corporate meetings, sales interactions, or customer support.

The purpose of this paper is to place summarization of spoken language in the context of general summarization research, describe its main challenges which are added on top of the already challenging area of written text summarization, describe past and current approaches and systems, and finally provide a tentative outlook on future directions in research and development of spoken language summarization systems.

2 A Brief History of Automatic Summarization

The first attempts to build systems that can summarize written texts automatically were made as early as the 1950s. In his seminal paper, (Luhn,

1958) constructed a summarizer to create abstracts from texts by selecting sentences that have a high number of significant words in close proximity to each other. Significance here means that words are neither too frequent (those words typically include closed class words such as articles, pronouns, modals, conjunctions etc.) nor too infrequent. About a decade later, (Edmundson, 1969) described an approach which expands on Luhn's earlier work in that it also takes into account genre-specific properties of the text, such as sentence location, words in titles, and certain cue or stigma phrases (to enhance or reduce sentence weight). Further, Edmundson also evaluated the quality of the automatic abstracts by comparing them to a human generated "gold standard". Most extract generation systems to date, even if they may employ more sophisticated techniques and may be more solidly grounded in statistics and theories of machine learning (such as (Kupiec, Pedersen, and Chen, 1995)), use these basic ideas of this early work: (1) determine significant, important words in the text; (2) select a set of features related to the text genre; (3) compute a score for each extraction segment (typically: a sentence, phrase, or paragraph) based on its relevance and the features' values; (4) extract the highest ranking segments in the order of the original text to form the abstract. If a reasonable number of (text,abstract)-pairs is available (e.g., in a collection of scientific articles and their author-generated abstracts), automatic summarizers can get good leverage from statistical training on these corpora. A good example is the system presented by (Banko, Mittal, and Witbrock, 2000) where headlines for newswire data are generated automatically.

Starting in the 1970s, summarization methods based on concepts in Artificial Intelligence became more *en vogue*. Unlike the approaches characterized so far, researchers were now concerned with identifying the underlying concepts in the text, to be able to understand the text, and then, by means of this understanding, reducing the information to its core in an abstract information representation, to finally generate a coherent, meaningful, and representative summary of the original text. While there have been successes in this area for projects focused on very limited domains, such as VERBMOBIL (Reithinger et al., 2000) or TOPIC (Reimer and Hahn, 1988), the complexity of a reasonably accurate semantic analysis of unrestricted domain texts is still considered to be far beyond the reach of realistic working systems.

In the 1990s, as large volumes of textual information became available online, and as the World Wide Web emerged, there was a decisive return to the "old" ideas of statistical sentence extraction for abstract generation.

Algorithms had to be efficient and, in some instances, allow for automatic training on corpora. Furthermore, as more summarization systems were developed by both academia and industry, the call for "objective evaluation" became stronger and eventually resulted in the first global evaluation of text summarization systems (SUMMAC) in 1998 (Mani et al., 1998). Recently, a new series of evaluations has been incepted in the framework of the DUC (Document Understanding Conference) (NIST, 2001). Some interesting outcomes of these evaluations were that (1) it turns out to be very hard to establish an "ideal" summary, since individual perspectives on what is relevant in a text may vary widely; (2) systems with quite different architectures and basic design may have similar performance scores overall but divergent scores on particular texts/tasks; (3) some texts are generally (much) easier to summarize than others but the reasons for this are so far barely explored; (4) due to the rather large human disagreement on relevance (and ideal summaries), it is often not too hard to automatically create summaries which are as similar to an ideal summary of a particular human annotator as the latter is similar to other human summaries.[1]

As significant advances in automatic speech recognition (ASR) were made in the 1990s, for the first time some researchers looked into the question of spoken language summarization. Since this is the topic of this paper, we will provide a brief overview in the following but a much more detailed account in a later section (section 5). The first applications in spoken language summarization were built in the context of speech-to-speech translation systems such as VERBMOBIL (Reithinger et al., 2000). Since these systems all operated in very restricted domains, a limited text understanding approach was feasible and allowed, in the case of VERBMOBIL, the generation of abstracts in multiple languages from a single knowledge representation format. Then, in the context of the DARPA Broadcast News workshops and TREC's Spoken Document Retrieval track, several systems were developed that enable the browsing, indexing, and retrieving of audio recordings, sometimes along with summarization of the contents, based on either human transcription or automatic transcription by ASR technology (Waibel, Bett, and Finke, 1998; Valenza et al., 1999; Whittaker et al., 1999). Other research on the other hand focused on the acoustic signal and made use of a variety of prosodic features to enable quick skimming/browsing (Arons, 1997) or extraction of passages which are prosodically emphasized (Chen and Withgott, 1992). Our own research progressed one step further

[1] If the last observation is good news or bad news, is left for the reader to decide.

and looked at summarization of spoken dialogues, conversations of two or more parties (Zechner, 2001).

3 Dimensions of Summarization

There are several dimensions which have to be considered when talking about summarization, such as the following:

- extracts vs. abstracts: While extracts are created by pure extraction of pieces of the original text (mostly: sentences, paragraphs, or clauses, sometimes keywords and/or keyphrases), abstracts are *generated* from some sort of semantic representation which reflects the logical structure of the text: the former can be done with entirely statistical methods (possibly enhanced with some linguistic knowledge), the latter requires not only a "deep" understanding of the text but also a generation component which produces intelligible text from the formal representation.

- indicative vs. informative: Indicative summaries are meant to give the user a rough idea about the main points of a text; these are typically used for tasks such as text classification or information retrieval; informative summaries should represent the most relevant information in a text and be able to serve as "surrogates" for the complete text.

- generic vs. query-driven: In the generic case, the summary should provide an unbiased view of the most relevant information in a text, if it is a query-driven summary, it should reflect the specific interests of this user by focusing on the query.

- single vs. multiple documents: Is there one text or several sources to summarize simultaneously? Multi-document summarization usually requires a much higher compression rate, along with a need for elimination of redundant information (Goldstein et al., 2000; Radev, Jing, and Budzikowska, 2000).

- background vs. just-the-news: In some instances, summarizers might have to be able to distinguish between these two kinds of information (specifically relevant for newswire data), e.g., to alert users to events which have not been reported in previous updates.

- single vs. multiple topics: Most short newswire articles (and research papers) will be mono-topical; however, there are many texts where this simplifying assumption does not hold and for which methods have to be established to reflect the multi-topicality in the summary.

- single vs. multiple speakers: The majority of text documents summarized will have a single speaker or writer; however, there are also interviews, discussions, conversations etc. where the information is distributed among multiple participants and sometimes is constructed by their interaction (e.g., by a question-answer pair).

- text-only vs. multi-modal: Summarization research so far almost exclusively focused on the written domain; in recent years, several research groups have started to explore how to summarize multi-modal and multi-media input (Waibel, Bett, and Finke, 1998; Waibel et al., 2001; Hirschberg et al., 1999; Valenza et al., 1999).

- selecting sentences/clauses vs. condensing within sentences: There has been a recent surge of research on trainable systems which can reduce the information *within* a sentence or a clause, whereas the mainstream of summarization research clearly has been concerned with sentence (or clause, paragraph) selection only. While (Jing, 2000) uses information from syntactic parses, context, and corpus statistics, (Knight and Marcu, 2000) use a noisy-channel and a decision tree model based on aligned parse trees of parallel corpora of (Text, Abstract) pairs.

4 Main Challenges

The main challenges that have to be addressed in spoken language summarization, in addition to the challenges of written text summarization, can be summarized as follows:

- coping with speech disfluencies

- identifying the units for extraction

- maintaining cross-speaker coherence (in case of multi-party conversations)

- coping with speech recognition errors

In the following, we shall discuss the nature of these challenges and indicate which approaches one can take to address them.

4.1 Disfluency detection

The two main negative effects speech disfluencies have on summarization are that they (i) decrease the readability of the summary and (ii) increase its non-content noise. In particular for informal conversations, the percentage of disfluent words is quite high, typically around 15-25% of the total words spoken. An example of a highly disfluent sentence, where the removal of disfluencies would enhance readability and conciseness of a summary, is given here:

```
A : well I um I think we should
    discuss this you know with her
A': I think we should discuss this with her
```

Previous work on speech disfluency detection and removal has used various machine learning approaches (such as decision trees), whose input features are typically a combination of word or part-of-speech information and a set of prosodic features (such as stress, pitch, and pauses) (Heeman and Allen, 1999; Stolcke and Shriberg, 1996).

4.2 Sentence boundary detection

Unlike written texts, where punctuation or hypertext markers indicate sentence boundaries, spoken language is generated as a sequence of streams of words, where pauses (silences between words) do not always match linguistically meaningful segments: a speaker can pause in the middle of a sentence or even a phrase, or, on the other hand, might not pause at all after the end of a sentence or a clause. If an audio stream is segmented into smaller units (e.g., *speaker turns*[2]) by means of using a silence heuristic, one speaker's turn may contain multiple sentences, or, on the other hand, a speaker's sentence might span more than one turn, as demonstrated in the following example:

```
1 A: That's true / I suggest
2 A: you talk to him /
```

The main problem for a summarizer would thus be (i) the lack of coherence and readability of the output because of incomplete sentences and (ii)

[2] A speaker turn is a contiguous part of a recording where one speaker is active.

extraneous information due to extracted units consisting of more than one sentence. Past work in automatic sentence segmentation used approaches such as language models, decision trees, or Hidden Markov Models, using textual and prosodic information (Stolcke, 1997; Heeman and Allen, 1999).

4.3 Distributed information

If we have multi-party conversations as opposed to monologues, sometimes the crucial information is found in a sequence of sentences from several speakers — the prototypical case being a question-answer pair. If the summarizer were to extract only the question or only the answer, the lack of the corresponding answer or question would often cause a severe reduction of coherence in the summary. In some cases, either the question or the answer is very short and does not contain any words with high relevance, resulting in a very small relevance weight within an automatic summarizer, e.g.:

```
A: Are you inviting all of your friends?
B: Yes.
```

In order not to lose these short sentences at a later stage, when only the most relevant sentences are extracted, one needs to identify matching question-answer pairs ahead of time, so that the summarizer can output these matching pairs during summary generation. We described an approach to cross-speaker information linking in (Zechner and Lavie, 2001). We used a decision tree to identify question speech acts first, and then used a set of trainable heuristics to determine the corresponding answers. A user study showed that while automatic question-answer linking may not increase the information content of a summary, it does increase its (local) coherence significantly.

4.4 Speech recognition errors

If there is no human generated transcription of an audio document available, the summarizer has to rely on an automatically generated transcription by a speech recognizer. Depending on the corpus, word error rates can typically range anywhere between 10% and 40%, the former number being applicable to more formal texts and clean channel conditions, the latter to more informal conversations with rather noisy channel conditions. We have shown in previous work that we can use speech recognizer confidence scores to (i) reduce the word error rate within the summary and (ii) increase the summary accuracy (Zechner and Waibel, 2000b).

5 Past Approaches

5.1 Summarization of spoken language in restricted domains

During the past decade, there has been significant progress in the area of closed domain spoken dialogue translation and understanding, even with automatic speech recognition input. Two examples of systems being developed in that time frame are JANUS (Lavie et al., 1997) and VERBMOBIL (Wahlster, 1993).

In that context, several spoken dialogue summarization systems were developed, whose goal it was to capture the essence of the task based dialogues at hand. The MIMI System (Kameyama and Arima, 1994; Kameyama, Kawai, and Arima, 1996) dealt with the travel reservation domain and used a cascade of finite state pattern recognizers to find the desired information.

Within VERBMOBIL, a more knowledge-rich approach was used (Alexandersson and Poller, 1998; Reithinger et al., 2000). The domain here is travel planning and negotiation of a trip. In addition to finite state transducers for content extraction and statistical dialogue act recognition, they also use a dialogue processor and a summary generator which have access to a world knowledge database, a domain model, and a semantic database. The abstract representations built by this summarizer allow for summary generation in multiple languages.

5.2 Summarization of spoken news

Within the context of the TREC spoken document retrieval (SDR) conferences (Garofolo et al., 1997; Garofolo et al., 1999) as well as the recent DARPA Broadcast News workshops, a number of research groups have been developing multi-media browsing tools for text, audio, and video data, which should facilitate the access to news data, combining different modalities.

(Hirschberg et al., 1999; Whittaker et al., 1999) present a system that supports local navigation for browsing and information extraction from acoustic databases, using speech recognizer transcripts in tandem with the original audio recording. While their interface helped users in the tasks of relevance ranking and fact-finding, it was less helpful in the creating of summaries, partly due to imperfect speech recognition.

Valenza et al. (1999) present an audio summarization system which combines acoustic confidence scores with relevance scores to obtain more accurate and reliable summaries. An evaluation showed that human judges prefer summaries with a compression rate of about 15% (30 words per minute at a speaking rate of about 200 words per minute), and that the summary word error rate was significantly smaller than the word error rate for the full transcript.

Hori and Furui (2000) use salience features in combination with a language model to reduce Japanese broadcast news captions by about 30-40% while keeping the meaning of about 72% of all sentences in the test set.

5.3 Prosody-based emphasis detection in spoken audio

While most approaches to summarizing of acoustic data rely on the word information (provided by a human or ASR transcript), there have been attempts to generate summaries based on emphasized regions in a discourse, using only prosodic features. Chen and Withgott (1992) train a Hidden Markov Model on transcriptions of spontaneous speech, labeled for different degrees of emphasis by a panel of listeners. Their "audio summaries" on an unseen (but rather small) test set receive a remarkably good agreement with human annotators ($\kappa > 0.5$). Stifelman (1995) uses a pitch based emphasis detection algorithm developed by Arons (1994) to find emphasized passages in a 13 minute discourse. In her analysis, she finds good agreement between these emphasized regions and the beginnings of manually marked discourse segments (in the framework of Grosz and Sidner (1986)). Although these are promising results, being suggestive of the role of prosody for determining emphasis, relevance, or salience in spoken language, further research needs to be done to explore these approaches in more depth and to also look into combining them with more traditional, text-based summarization methods.

5.4 Spoken dialogue summarization in unrestricted domains

Waibel, Bett, and Finke (1998) report results of their summarizer on automatically transcribed SWITCHBOARD data (Godfrey, Holliman, and McDaniel, 1992), the word error rate being about 30%. Their implementation used an algorithm inspired by maximum marginal relevance (MMR) (Carbonell, Geng, and Goldstein, 1997), but they did not address any dialogue or speech related issues in their summarizer. In a question-answer test with summaries of five dialogues, subjects could identify most of the key concepts using a summary size of only five turns. These results varied widely across five different dialogues tested in this experiment (between 20% and 90% accuracy).

Our own work (Zechner and Waibel, 2000a; Zechner, 2001) presented a summarization system (DIASUMM) for spoken dialogues in unrestricted domains for the first time. The DIASUMM system addresses the issues of disfluency detection and removal, sentence boundary detection, as well as cross-speaker information linking and ASR word error rate reduction. The components of the DIASUMM system were trained on a large corpus of disfluency annotated conversations (LDC, 1999) and tested on four different genres of spoken dialogues. We were able to show that for more informal genres of conversations, the DIASUMM system outperformed two baselines significantly (LEAD[3], MMR).

6 Conclusion and Outlook

During the past decade, there has been an increasing interest in summarization of dialogues and audio documents of various kinds. As a result, a number of different methods and approaches have been proposed and a variety of systems have been built that can perform different summarization tasks on spoken language input.

We think that in the near future, as the amount of digitized speech available on-line will rise substantially, the research into developing robust summarization technology for this genre will have to involve the following aspects:

- Robust and improved speech recognition: ASR will have to be performed across a wide range of channel conditions, sometimes with substantial noise, cross-talk and other hard parameters. Current state-of-the-art speech recognizers still exhibit fairly high average word error rates of up to 40% for challenging genres such as multi-party meetings in noisy environments with participants using rather conversational speech.

- Integration of prosodic and word-based information: While so far systems rely on either kind of information (more or less) exclusively,

[3]LEAD baseline: Extract the beginning of a text segment. This is a usually very successful strategy for summarization in newswire domains.

we think that it will be essential to combine acoustic and word-level information so that they can complement each other for the benefit of a variety of summarization sub-tasks, such as disfluency detection, sentence boundary identification, topic segmentation, and emphasis detection.

- Development of meaningful annotation and evaluation procedures to facilitate comparative system evaluations: While in the field of written text summarization, methods to annotate, evaluate, and compare different summarization approaches and systems have been developed for some years now, there is a complete lack of uniform standards in the area of spoken language summarization. While some methods may be transferable, others may have to be changed or added to accommodate the difference of the input source (i.e., spoken vs. written). We think also that in addition to automatic evaluations, user studies will have to be performed, where users can use both textual and acoustic information of audio summaries to perform a specific task.

Finally, we believe that the role of summarization of spoken language in the future will be found in supporting a continuum of different ways to aide a user to meet his or her information needs, ranging from question-answering over document or passage retrieval to data mining and information extraction.

References

Alexandersson, Jan and Peter Poller. 1998. Towards multilingual protocol generation for spontaneous speech dialogues. In *Proceedings of the INLG-98, Niagara-on-the-lake, Canada, August*.

Arons, Barry. 1994. Pitch-based emphasis detection for segmenting speech. In *Proceedings of the ICSLP-94*, pages 1931–1934.

Arons, Barry. 1997. SpeechSkimmer: A system for interactively skimming recorded speech. *ACM Transactions on Computer Human Interaction*, 4(1):3–38, March.

Banko, Michele, Vibhu O. Mittal, and Michael J. Witbrock. 2000. Headline generation based on statistical translation. In *Proceedings of the 38th Conference of the Association for Computational Linguistics, Hongkong, China, October*, pages 318–325.

Carbonell, Jaime, Yibing Geng, and Jade Goldstein. 1997. Automated query-relevant summarization and diversity-based reranking. In *Proceedings of the IJCAI-97 workshop on AI and digital libraries, Nagoya, Japan*.

Chen, Francine R. and Margaret Withgott. 1992. The use of emphasis to automatically summarize a spoken discourse. In *Proceedings of the ICASSP-92*, pages 229–332.

Edmundson, H. P. 1969. New methods in automatic extracting. *Journal of the ACM*, 16(2):264–285, April.

Garofolo, John S., Ellen M. Voorhees, Cedric G. P. Auzanne, and Vincent M. Stanford. 1999. Spoken document retrieval: 1998 evaluation and investigation of new metrics. In *Proceedings of the ESCA workshop: Accessing information in spoken audio*, pages 1–7. Cambridge, UK, April.

Garofolo, John S., Ellen M. Voorhees, Vincent M. Stanford, and Karen Sparck Jones. 1997. TREC-6 1997 spoken document retrieval track overview and results. In *Proceedings of the 1997 TREC-6 Conference, Gaithersburg, MD, November*, pages 83–91.

Godfrey, J. J., E. C. Holliman, and J. McDaniel. 1992. SWITCHBOARD: telephone speech corpus for research and development. In *Proceedings of the ICASSP-92*, volume 1, pages 517–520.

Goldstein, Jade, Vibhu Mittal, Jaime Carbonell, and Mark Kantrowitz. 2000. Multi-document summarization by sentence extraction. In *Proceedings of the NAACL-ANLP-2000 Workshop on Automatic Summarization, Seattle, WA, April*.

Grosz, Barbara J. and Candace L. Sidner. 1986. Attention, intentions, and the structure of discourse. *Computational Linguistics*, 12(3):175–204.

Heeman, Peter A. and James F. Allen. 1999. Speech repairs, intonational phrases, and discourse markers: Modeling speakers' utterances in spoken dialogue. *Computational Linguistics*, 25(4):527–571, December.

Hirschberg, Julia, Steve Whittaker, Don Hindle, Fernando Pereira, and Amit Singhal. 1999. Finding information in audio: A new paradigm for audio browsing/retrieval. In *Proceedings of the ESCA workshop: Accessing information in spoken audio*, pages 117–122. Cambridge, UK, April.

Hori, Chiori and Sadaoki Furui. 2000. Automatic speech summarization based on word significance and linguistic likelihood. In *Proceedings of ICASSP-00, Instanbul, Turkey, June*, pages 1579–1582.

Jing, Hongyan. 2000. Sentence reduction for automatic text summarization. In *Proceedings of ANLP-NAACL-2000, Seattle, WA, May*, pages 310–315.

Kameyama, M. and I. Arima. 1994. Coping with aboutness complexity in information extraction from spoken dialogues. In *Proceedings of the ICSLP 94, Yokohama, Japan*, pages 87–90.

Kameyama, Megumi, Goh Kawai, and Isao Arima. 1996. A real-time system for summarizing human-human spontaneous spoken dialogues. In *Proceedings of the ICSLP-96*, pages 681–684.

Knight, Kevin and Daniel Marcu. 2000. Statistics-based summarization — step one: Sentence compression. In *Proceedings of the 17th National Conference of the AAAI*.

Kupiec, J., J. Pedersen, and F. Chen. 1995. A trainable document summarizer. In *Proceedings of the 18th ACM-SIGIR Conference*, pages 68–73.

Lavie, Alon, Alex Waibel, Lori Levin, Michael Finke, Donna Gates, Marsal Gavaldà, Torsten Zeppenfeld, and Puming Zhan. 1997. Janus III: Speech-to-speech translation in multiple languages. In *IEEE International Conference on Acoustics, Speech and Signal Processing, Munich, 1997*.

LDC, Linguistic Data Consortium. 1999. Treebank-3: CD-ROM containing databases of disfluency annotated Switchboard transcripts (LDC99T42).

Luhn, H. P. 1958. The automatic creation of literature abstracts. *IBM Journal of Research and Development*, 2(2):159–165.

Mani, Inderjeet, David House, Gary Klein, Lynette Hirschman, Leo Obrst, Therese Firmin, Michael Chrzanowski, and Beth Sundheim. 1998. The TIPSTER SUMMAC text summarization evaluation. Mitre Technical Report MTR 98W0000138, October 1998.

NIST. 2001. Document Understanding Conference (DUC) 2001. http://www-nlpir.nist.gov/projects/duc/.

Radev, Dragomir R., Hongyan Jing, and Malgorzata Budzikowska. 2000. Centroid-based summarization of multiple documents: sentence extraction, utility-based evaluation, and user studies. In *Proceedings of the NAACL-ANLP-2000 Workshop on Automatic Summarization, Seattle, WA, April*, pages 21–30.

Reimer, U. and U. Hahn. 1988. Text condensation as knowledge base abstraction. In *Proceedings of the 4th Conference on Artificial Intelligence Applications, San Diego, CA*, pages 338–344.

Reithinger, Norbert, Michael Kipp, Ralf Engel, and Jan Alexandersson. 2000. Summarizing multilingual spoken negotiation dialogues. In *Proceedings of the 38th Conference of the Association for Computational Linguistics, Hongkong, China, October*, pages 310–317.

Stifelman, Lisa J. 1995. A discourse analysis approach to structured speech. In *AAAI-95 Spring Sympoisium on Empirical Methods in Discourse Interpretation and Generation, Stanford, CA, March*.

Stolcke, Andreas. 1997. Modeling linguistic segment and turn boundaries for N-best rescoring of spontaneous speech. In *Proceedings of EUROSPEECH-97, Rhodes, Greece*.

Stolcke, Andreas and Elizabeth Shriberg. 1996. Statistical language modeling for speech disfluencies. In *Proceedings of the ICASSP-96, Atlanta, GA, May*.

Valenza, Robin, Tony Robinson, Marianne Hickey, and Roger Tucker. 1999. Summarisation of spoken audio through information extraction. In *Proceedings of the ESCA workshop: Accessing information in spoken audio*, pages 111–116. Cambridge, UK, April.

Wahlster, Wolfgang. 1993. Verbmobil — translation of face-to-face dialogs. In *Proceedings of MT Summit IV, Kobe, Japan*.

Waibel, Alex, Michael Bett, and Michael Finke. 1998. Meeting browser: Tracking and summarizing meetings. In *Proceedings of the DARPA Broadcast News Workshop*.

Waibel, Alex, Michael Bett, Florian Metze, Klaus Ries, Thomas Schaaf, Tanja Schultz, Hagen Soltau, Hua Yu, and Klaus Zechner. 2001. Advances in automatic meeting record creation and access. In *Proceedings of ICASSP-2001, Salt Lake City, UT, May*.

Whittaker, Steve, Julia Hirschberg, John Choi, Don Hindle, Fernando Pereira, and Amit Singhal. 1999. SCAN: Designing and evaluating user interfaces to support retrieval from speech archives. In *Proceedings of the 22nd ACM-SIGIR International Conference on Research and Development in Information Retrieval, Berkeley, CA, August*, pages 26–33.

Zechner, Klaus. 2001. *Automatic Summarization of Spoken Dialogues in Unrestricted Domains*. Ph.D. thesis, Language Technologies Institute, School of Computer Science, Carnegie Mellon University, CMU-LTI-01-168, November.

Zechner, Klaus and Alon Lavie. 2001. Increasing the coherence of spoken dialogue summaries by cross-speaker information linking. In *Proceedings of the NAACL-01 Workshop on Automatic Summarization, Pittsburgh, PA, June*, pages 22–31.

Zechner, Klaus and Alex Waibel. 2000a. DIASUMM: Flexible summarization of spontaneous dialogues in unrestricted domains. In *Proceedings of COLING-2000, Saarbrücken, Germany, July/August*, pages 968–974.

Zechner, Klaus and Alex Waibel. 2000b. Minimizing word error rate in textual summaries of spoken language. In *Proceedings of the First Meeting of the North American Chapter of the Association for Computational Linguistics, NAACL-2000, Seattle, WA, April/May*, pages 186–193.